Exploring Semantic Spaces for Detecting Clustering and Switching in Verbal Fluency

Özge Alaçam[1], Simeon Schüz[1], Martin Wegrzyn[2], Johanna Kißler[2] and Sina Zarrieß[1]
[1]Computational Linguistics, [2]Department of Psychology,
University of Bielefeld, Bielefeld, Germany
{oezge.alacam, simeon.schuez, martin.wegrzyn,
johanna.kissler, sina.zarriess}@uni-bielefeld.de

Abstract

In this work, we explore the fitness of various word/concept representations in analyzing an experimental verbal fluency dataset providing human responses to 10 different category enumeration tasks. Based on human annotations of so-called clusters and switches between sub-categories in the verbal fluency sequences, we analyze whether lexical semantic knowledge represented in word embedding spaces (GloVe, fastText, ConceptNet, BERT) is suitable for detecting these conceptual clusters and switches within and across different categories. Our results indicate that ConceptNet embeddings, a distributional semantics method enriched with taxonomical relations, outperforms other semantic representations by a large margin. Moreover, category-specific analysis suggests that individual thresholds per category are more suited for the analysis of clustering and switching in particular embedding sub-space instead of a one-fits-all cross-category solution. The results point to interesting directions for future work on probing word embedding models on the verbal fluency task.

1 Introduction

The intrinsic evaluation of lexical knowledge represented in word embeddings has been of long-standing interest in distributional semantics (Levy et al., 2015; Hill et al., 2015), and remains an important topic in work on interpreting large-scale black-box language models (Pezzelle et al., 2021; Vulić et al., 2020; Bommasani et al., 2020). While pre-trained contextualized word representations have recently been evaluated in many novel, (psycho-) linguistically motivated probing tasks (Belinkov and Glass, 2019; Ettinger, 2020; Finlayson et al., 2021), the assessment of lexical semantics in word embeddings still commonly focuses on traditional benchmarks of human similarity annotations (Hill et al., 2015), datasets of analogies (Drozd et al., 2016) or taxonomic relations such as hypernymy (Baroni and Lenci, 2011; Glavaš and Vulić, 2018).

In this paper, we conduct an evaluation of word embeddings on the so-called verbal fluency task (Shao et al., 2014), where participants are asked to enumerate as many different words for a given category as possible within a given time interval (often 60 seconds), see Figure 1 for an example response to the category *hobby*. The resulting production data are a rich source of participants' lexical-conceptual knowledge and typically show an interesting clustering-switching pattern where consecutive words either relate to the same sub-category (*handball, football,horseback riding* as a cluster for "sports" in Figure 1) or switch between sub-categories (*playing the guitar, model railway* as a switch from "music" to "playing" in Figure 1). Verbal fluency is a very well-known and widely used cognitive performance test used in psychology, neuro- and psycholinguistics where robust and automatic methods for analyzing clustering-switching patterns would be highly welcome (Kim et al., 2019). Yet, to date, the analysis of verbal fluency data received little attention in research on computational semantics and word embeddings (Pauselli et al., 2018; Linz et al., 2017; Pakhomov and Hemmy, 2014). The study by Linz et al. (2017) constitutes a noticeable exception, but is restricted to verbal fluency responses to a single category ("animals") and does not rely on human annota-

1	Malen *[painting]*	Schaffend *[creating]*
2	Lesen *[reading]*	Lyrik *[lyric]*
3	Schreiben *[writing]*	Lyrik *[lyric]*
4	Handball *[handball]*	Sport *[sport]*
5	Fußball *[football]*	Sport *[sport]*
6	Reiten *[horseback riding]*	Sport *[sport]*
7	Musik *[music]*	Musik *[music]*
8	Gitarre spielen *[playing the guitar]*	Musik *[music]*
9	Modelleisenbahn *[model railway]*	Spielen *[playing]*
10	Modellflugzeug *[model airplane]*	Spielen *[playing]*
11	Sammeln *[collecting]*	Sammeln *[collecting]*
12	Stricken *[knitting]*	Schaffend *[creating]*
13	Sticken *[embroidery]*	Schaffend *[creating]*
14	Nähen *[sewing]*	Schaffend *[creating]*

Figure 1: Elicited word sequence and annotated sub-categories from the hobbies domain

Proceedings of the 29th International Conference on Computational Linguistics, pages 178–191
October 12–17, 2022.

tions of clusters and switches.

We base our study on a recently collected dataset of German semantic verbal fluency responses that provides a much wider range of categories than previous studies and, additionally, has been analyzed in terms of clustering-switching patterns by human judges.[1] Importantly, our verbal fluency-based evaluation of lexical knowledge in embedding spaces rests on human production and judgement data such that it may display different aspects of taxonomical-conceptual knowledge as compared to existing benchmarks for lexical relation prediction derived from standard lexical resources (Baroni and Lenci, 2011; Glavaš and Vulić, 2018). For instance, the sub-categories for the "hobby" category illustrated in Figure 1 could not be easily retrieved from WordNet (Fellbaum, 2010) which, curiously, lists "speleology" as the only direct hyponym for the most common synset of "hobby".

In the following, we compare different word embeddings for German (BERT-base, GloVe, Fasttext, ConceptNet), investigating to what extent distances in sub-spaces for categories like *animals, body parts, clothes* reflect conceptual switches found in verbal fluency data and to what extent it is possible to derive generic, cross-category distance thresholds for the accurate detection of clusters and switches. Our results indicate that ConceptNet embeddings (Speer et al., 2017), a distributional semantics method enriched with taxonomical relations, outperforms other semantic representations by a large margin. On the other hand, merely taxonomical relations (as represented in GermaNet Hamp and Feldweg (1997)) have significant but weak correlations indicating that they are useful as accompanying modalities to word embeddings. Moreover, category-specific analysis suggests that individual thresholds per category are more suited for the analysis of clustering and switching in particular embedding sub-spaces instead of a one-fits-all cross-category solution. A final experiment using simple clustering algorithms further corroborates the findings of the switching analyses. Overall, the results point to interesting directions for future work on probing word embedding models on the verbal fluency task.

[1]The dataset has not been released for ethical reasons, but can be obtained under restricted conditions, upon request.

2 Background

The verbal fluency task is a very well-known neuropsychological test that is used in clinical contexts for diagnosing, e.g., neurodegenerative diseases as well as in research on the cognitive processes underlying lexical knowledge, access, retrieval and executive control (Shao et al., 2014). Participants' verbal fluency responses are typically scored in terms of the number of correct words produced for the category, whereas more fine-grained analyses measure the amount of switching and clustering in the word sequence (Troyer et al., 1997). The semantic analysis of verbal fluency is commonly addressed by the manually defined subcategories for "animals" established in Troyer et al. (1997)'s study, and extending these to other categories and languages is a notorious challenge in psychology (Kim et al., 2019). While there has been a lot of interest in psychology in using word embeddings for scoring semantic fluency (Benigni et al., 2021; Qiu and Johns, 2021; Kim et al., 2019; Paula et al., 2018; Linz et al., 2017), the task entered the radars of the NLP community only recently. One potential reason is that switching and clustering literature are restricted to an extremely limited number of categories (such as *animals*, *groceries*), although there are standardized tools that slightly extend the list of the categories — e.g. the popular RWT (Regensburger Wortflüssigkeitstest) (Aschenbrenner et al., 2000) includes five categories (*animals*, *hobbies*, *occupations*, *groceries* and *first names*).

Quantitative analyses of verbal fluency data have shown that enumeration speed and diversity within a category are very category-dependent and that categories can be more or less easy to enumerate. For automatic methods that measure clustering and switching, a first key step is to define appropriate thresholds that pinpoint switch boundaries for subcategory changes. Kim et al. (2019) investigate different ways of automatically scoring semantic fluency in English and Korean. Using a traditional word2vec model (Mikolov et al., 2013), they predict categorical switches in collected sequences if the predicted similarity between adjacent items drops below a defined threshold, similar to the approach in (Linz et al., 2017). Complementary to this, the authors propose a model which aligns words from fluency sequences with Wikipedia articles and predicts categorical switches when the intersection of articles linked to adjacent words drops below a certain threshold. Despite well-known dif-

ferences between verbal fluency categories, the robustness and quality of these threshold-based methods across categories has, to the best of our knowledge, not yet been analyzed. Moreover, from an NLP perspective, traditional word2vec embeddings can be expected to achieve a lower performance in analyzing fine-grained conceptual relations as compared to various more recent embedding methods that capture global word distributions as in GloVe (Pennington et al., 2014), subword representations as in fastText (Bojanowski et al., 2017) or integrated taxonomical knowledge as ConceptNET Numberbatch (Egozi et al., 2011). Finally, contextualized embeddings from transformer language models such as BERT (Devlin et al., 2019) constitute to be an obvious method to explore. However, to be used in such enumeration task, this dynamic embedding method needs to be transformed into static embeddings as detailed in Section 3.3. In brief, the main contribution of this paper is to explore (i) state-of-the-art word embeddings for data collected in an ongoing verbal fluency study in wide categorical variety and (ii) the options for automatic scoring mechanisms for this broad range of categories. To our knowledge, this is the first (NLP-powered) study that systematically analyzes verbal fluency task across such categorical variety through various semantic representations and evaluation metrics together.

3 Experiments

3.1 Data

In this section, we describe the data collection, annotation and cleaning protocols. Detailed information is provided in Appendix 7.1

Participants. 125 participants originally attended the study, and 114 of them completed it. After cleaning, 100 participants are included in the following analysis (age: 18-63 ($mean$ = 26), $gender$: 87 *female*, 10 *male*, 3 *non-binary*).

Semantic Categories. The initial dataset contains 24 conceptual categories. However, not all of them resulted in sufficient data for statistical analysis. Second, some categories like *amphibians* and *precious stones* elicit a considerable amount of rare words which do not exist in the vocabulary of the methods used here. Furthermore, some categories are very subjective and less related to linguistic lexical knowledge as, e.g., *first names*). Therefore based on qualitative and descriptive analysis, we narrow the 24 categories down to those that have at least 75 words produced by the probands, which are available in all embeddings' vocabulary list, and with a minimum average of 5 words per annotated subcategory. This leaves us with the following 10 categories: *occupations, groceries, hobbies, animals, weapons, vessels, fabrics, countries, clothes, body parts* and *insects*. The entire list can be found in Appendix 7.2.

Subcategory Annotations The words in verbal fluency sequences have been manually annotated with their subcategories (e.g. *pets*, *birds*, *jungle* animals for the *animal* category) by five paid, trained annotators, each annotating 4-5 of the 24 categories. Based on this annotation, we are able to determine *switches* (positions where the left and right word have a different subcategory and *clusters* (sequences of words with the same category).

Data Cleaning. We remove sequences that contain less then 5 items, resulting in 960 sequences in total. The words in the sequences were processed using off-the-shelf NLP text processing tools like; SpaCy Lemmatizer[2], Compound Splitter[3] and Spell Checker[4] for German. Compound words are generally common in German and the vocabulary used by participants also frequently contains compound words such as "Klavierspielen" (piano playing), "Krankenpfleger" (health nurse), "Fahrradfahren" (bike riding). Unfortunately, many of the compounds do not exist in the vocabulary of GloVe, ConceptNet, or GermaNet whereas fastText and BERT embeddings can deal with out-of-vocabulary tokens due to their sub-word tokenization method. In order to address this discrepancy for the non-subword methods, the Python compound splitter package has been used for the words not found in the vocabulary following the lemmatization and spell-check. As a result, the compound vector would be the average of the part vectors.

Table 1 presents basic statistics for word counts and sub-category switch counts observed in the sequences within each category and across categories (as *global*) following the method used by Kim et al. (2019). This overview highlights the differences in the characteristics of the categories: participants enumerated almost 20 items on average for the *animals* and *countries*, and around or below 10 items for *fabrics*, *insects*, and *vessels*. Correspondingly,

[2] https://spacy.io/models/de
[3] https://github.com/dtuggener/CharSplit
[4] https://pypi.org/project/pyspellchecker/

switch counts for *animals* and *hobbies* are significantly higher as compared to categories which are less easy to enumerate.

3.2 Methods

We now introduce our automatic switch detection methods, that we will evaluate on the human subcategory annotations. The goal is to investigate whether it is possible to determine a "one-fits-all" metric that can generalize across various semantic categories and to further explore category-dependent characteristics that cause deviation from the overall pattern.

In addition to comparing human-annotations with the word/concept embedding methods (GloVe, fastText, ConceptNet and BERT-base), we further investigate how mere taxonomic relations (by employing GermaNet) perform on switch detection (i) as a standalone method (Section 4.3) and (ii) as a complementary source of information, combined with embedding-based decisions (Section 4.4).

We utilize several metrics in order to test the correlations between human annotations and embedding representations. First, we compare the the number of switches determined by the human annotators against the aggregated similarity score calculated for the pairs in the sequences (for each method, Section 4.1). We consider several parameters (*mean*, *maximum*, *minimum* and *standard deviation*) for our aggregation method. Since *mean* values exhibit the highest correlation scores, we select them as the suitable scoring metric for reporting.

Next, we try to detect switch boundaries (subcategory changes) in sequences, using the similarity scores between word pairs in the sequences. To decide whether a given word pair marks a subcategory switch, we apply the threshold cut-off methods described in Kim et al. (2019). We test two threshold variations: (i) Median threshold and (ii) 25-Percentile (25P) threshold. A switch boundary is marked where the cosine similarity between two adjacent words falls below the respective threshold.

For the individual embedding methods, the binary threshold-based decisions whether pairs of words mark switch boundaries or belong to the same subcategory are then compared against the human annotations using Cohen Kappas and Chi-square statistics (using the *scipy* package[5], Section 4.2). The median and 25P thresholds are calculated per category as well as globally, by taking all similarity values in the entire data into account. The entire list of calculated thresholds can be found in Appendix 7.3.

We complement our embedding-based analyses with GermaNet (Hamp and Feldweg, 1997), a lexical-semantic network for German that allows for a rule-based, explainable analysis of the switch boundaries. GermaNet groups nouns, verbs, and adjectives into synsets and links these synsets with lexical semantic relations (containing a total of 205K lexical units in 159K synsets). Using the Python API for GermaNet (*germanetpy*[6]), we extract the lexical units and synsets for the word pairs given their category.

We explore the following metrics for scoring similarity between word pairs based on synset relations: (i) shortest path distance, (ii) path-based (PB) similarity and (iii) information content (IC) based similarity (Resnik, 1999; Leacock and Chodorow, 1998). The details of these metrics can be found in the GermaNet website with a source code[7]. Path-based relatedness measures compute the semantic relatedness between two concepts based on the shortest path between two synsets in the hypernym relation. However, quantifying semantic distances merely based on length in the hypernym relation is intuitively not a flawless concept (Jiang and Conrath, 1997). The IC-based metric combines the structural information in the hypernym relation with the word frequencies (GermaNet raw frequency lists). The relatedness of two synsets is measured in terms of the information content of the least common synset that is a hypernym to both synsets. The word frequencies are used to compute the information content, which scores concepts from specific to general. If a very specific synset is compared to a very general one, the relatedness score will be low. This makes IC-based measures more suited for the similarity annotations. The formulas of these measures are available in Gurevych and Niederlich (2005). In the following, we focus on the IC-based metric due to its superior performance on our data. Detailed scores for all three metrics are provided in Appendix 7.5.

[5] https://docs.scipy.org/doc/scipy/reference/stats.html

[6] https://pypi.org/project/germanetpy/

[7] https://github.com/Germanet-sfs/germanetTutorials

Table 1: Basic statistics (Max, min, and average values of sequences and sub-category switches)

Categories	Word Count in a Sequence	Sub-category switch in a sequence	Total Word Count	Subcategory Count
animals	**Max: 30, Min: 5, Mean: 19.11**	Max: 14.0, Min: 1.0, Mean: 7.8	1659	22
body parts	**Max: 31 , Min: 5 , Mean: 18.2**	Max: 15.0, Min: 3.0 , Mean: 8.37	1527	8
clothes	Max: 24, Min: 7, Mean: 16.5	Max: 13.0, Min: 3.0, Mean: 8.14	1434	15
countries	**Max: 36, Min: 10, Mean: 18.5**	Max: 13.0, Min: .0, Mean: 4.6	1752	6
fabrics	Max: 17, Min: 5, Mean: 7.8	Max: 8.0, Min: .0, Mean: 3.1	537	15
groceries	Max: 25, Min: 6, Mean: 16.6	**Max: 16.0, Min: 3.0, Mean: 9.3**,	1520	14
hobbies	Max: 22, Min: 5, Mean: 14.4	**Max: 15.0, Min: .0, Mean: 7.7**	1158	31
insects	Max: 17, Min: 5, Mean: 9.8	Max: 11.0, Min: 2.0, Mean: 6.4	773	14
occupations	Max: 17, Min: 5, Mean: 12.5	Max: 13.0, Min: 3.0, Mean: 8.3	964	19
vessels	Max: 17, Min: 5, Mean: 10.1	Max: 12.0, Min: 1.0, Mean: 5.9	753	9
Global	Max: 36, Min: 5, Mean: 13.9	Max: 14, Min: 0, Mean: 6.8	12077	153

3.3 Semantic Space Representations

As introduced before, we investigate GloVe (1.31M vocab, 300 dimensional)[8] and fastText (65B tokens, 20M vocab, 300 dim.)[9] as general static word representations for German. As a third method, we test the ConceptNET Numberbatch word embeddings, which are enriched by ConceptNet taxonomic relations (594K vocab, 300 dim.) (Speer et al., 2017). Considering the task at hand, those relations might facilitate the enumeration.

Furthermore, we include BERT embeddings as one of the currently most popular models in NLP. Here, one potential challenge is that sequences of words in verbal fluency data differ substantially from the context that these models are trained for. Many layers of linguistic information like syntax or morphology that transformers learn to represent in their latent layers (Tenney et al., 2019) are not instrumental for this task. In this respect, verbal fluency data differs from most existing probing setups which prompt language models with "regular" linguistic inputs. Therefore, we convert contextualized BERT embeddings (2,350M tokens, 31K vocab, 512 dim.) to static word embeddings following the method explained in Bommasani et al. (2020). For this, we sample 20 sentences from the German Wikipedia 2 Corpus[10] for each item in our vocabulary, and compute their vectors using the *dbmdz/bert-base-german-cased* model[11]. After applying a pooling strategy, we end up with a static/single representations for each word.

[8] https://www.deepset.ai/german-word-embeddings

[9] https://fasttext.cc/docs/en/crawl-vectors.html

[10] https://github.com/GermanT5/wikipedia2corpus

[11] https://huggingface.co/dbmdz/bert-base-german-cased

4 Pairwise Switch Analysis

4.1 Switch Count - Similarity Score Correlations

The correlation between the switch counts per sequence (in total 960 sequences) and the mean cosine similarity score of the sequences is analyzed using the Pearson Correlation coefficient (scipy.stats.pearsonr[12]). Sequences with more sub-category switches are expected to have lower (mean) similarity scores. As illustrated in Table 2, the negative correlation is strong for only some of the categories such as *animals*, *countries*, *groceries*, and *insects*. The switch counts for GloVe and BERT embedding methods do not display convincing alignment with the human annotations, whereas fastText and ConceptNet embeddings are more in line with human decisions. This analysis shows that despite some significant and strong correlated categories, especially for the categories *clothes*, *fabrics*, *occupations* and *vessels*, no correlation has been observed indicating that the switch detection methods applied in the following section based on word embeddings might be challenging on these categories.

Table 2: Pearson Correlation Analysis Results on Total Switch Count and Mean Similarity Scores

Categories	GloVe	fastText	ConceptNet	BERT
animals	–.17, n.s.	**–.25, p.<.05**	**–.24, p.<.05**	–.07, n.s.
body parts	–.19, n.s.	.09, n.s.	**.23, p.<.05**	.30, p.<.01
clothes	.03, n.s.	–.147, n.s.	–.14, n.s.	.04, n.s.
countries	**–.43, p.<.01**	**–.39, p.<.01**	**–.40, p.<.01**	.02, n.s.
fabrics	–.11, n.s.	.01, n.s.	–.19, n.s.	.12, n.s.
groceries	**–.34, p.<.01**	**–.24, p. <.05**	**–.27, p.<.01**	–.08, n.s.
hobbies	–.144, n.s.	.03, n.s.	–.17, n.s.	.058, n.s.
insects	–.19, n.s.	**–.38, p.<.01**	**–.27, p.<.01**	–.13, n.s.
occupations	–.00, n.s.	–.16, n.s.	–.08, n.s.	.058, n.s.
vessels	.17, n.s.	–.01, n.s.	–.03, n.s.	.06, n.s.

[12] https://docs.scipy.org/doc/scipy/reference/stats.html

4.2 Embedding-based Switch Detection

As the previous metric returns mixed results, we continue our analysis by turning the continuous similarity scores into discrete switch boundaries.

Overall, switch detection based on 25-Percentile thresholds seems to achieve significant correlations with human annotations (Table 3). While the correlations are weak for GloVe and BERT embeddings, fastText and especially ConceptNet are correlated with the human annotation at various strengths. Categories like *animals*, *hobbies* and *countries* show stronger correlations, while less common categories like *vessels*, *fabrics*, *insects* achieve lower scores.

In addition to this, Table 4 presents the median thresholds, which are more conservative by design. The overall results confirm the category-dependent variations. ConceptNet scores are closer to human scores in 8 of 9 categories, showing strong-to-moderate correlations. Furthermore, the median thresholds seem to be aligned better with the human annotations than the 25-Percentile thresholds for this enumeration task. Similar to previous results, categories like *vessels*, and *fabrics* results in less alignment for this metric as well.

Global threshold vs. category threshold. In order to investigate how a global threshold compares to category-dependent ones, we test the alignment scores of both kinds of thresholds to human-annotations. On average, as illustrated in Figure 2, category-dependent threshold decisions demonstrates slightly better correlations with human annotations.The improvements are particularly evident with regard to some specific configurations: BERT ($avg = .02$, $max = .18$ in insects), ConceptNet ($avg = .01$, $max = .15$ in insects), fastText ($avg = .01$, $max = .11$ in body parts), GloVe ($avg = .02$, $max = .10$ in countries).

4.3 Taxonomy-based Switch Detection

Table 5 presents average shortest path and IC-based similarity scores per category. Based on these scores, median thresholds for each category are calculated following the threshold cut-off method explained in Section 3.2. In detecting sub-category switches, IC-based similarity shows alignment with the human annotations at various levels. The results indicate considerable differences between the categories. For example, whereas high correlations are observed for *hobbies*, *body parts*, *animals*, *occupations* and interestingly *insects*, the correlations are weaker for other categories like *clothes* or *fabrics*.

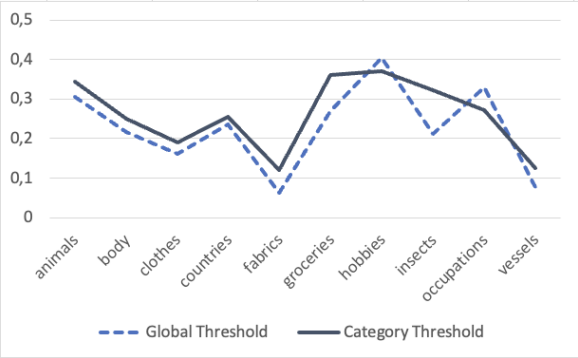

Figure 2: Correlations between the automated scores (Global- vs. Category-Thresholds) and human-annotations — averaged for all methods

4.4 Integrated Switch Detection

To investigate the influence of taxonomic relations in a more explicit way, we combined the decisions on subcategory switch from each embedding model with the decisions calculated using GermaNet IC-based scores (after converting to switch sequences using Median threshold). For each pair in a sequence, the combined metric predicts a switch if there is a switch detected in either of the sequences.

Figure 3 shows the correlation patterns (against human-annotated switch boundaries) across categories for each embedding method with (W+G) and without (W) GermaNet integration. Despite category-based significant differences between W and W+G conditions, it is difficult to conclude on an overall pattern that can explain the variations across categories. However, except ConceptNet ($avg = -.01$, $max = +.07$, higher improvement for *groceries*, *fabrics*), the other representation methods benefited from the inclusion of GermaNet relations especially for some categories; BERT ($avg = +.11$, $max = +.19$, high diff. in *body parts*, *vessels*, *fabrics*, *groceries*), fastText ($avg = +.08$, $max = +.12$, high diff. in fabrics), GloVe ($avg = +.014$, $max = +.23$, high diff. in *vessels*, *insects*, *fabrics*, *groceries*). Thus, the detection of switch boundaries for *vessels* and *fabrics* benefited most from the combined metric, which suggests that embedding models and lexical resources represent complementary aspects of lexical knowledge for these categories. Moreover, GloVe and BERT embeddings benefit more from GermaNet-informed scores (W+G) than fastText and ConceptNet, both for single categories and in the global evaluation (the last item in the graph).

Table 3: Interrater agreement between human annotations and 25-Percentile Thresholds for each embedding

Categories	Human-Annotated Data versus Embeddings' 25-Percentile Threshold			
	BERT	**GloVe**	**ConceptNet**	**fastText**
animals	κ: .07, Corr.: .08, p. <.01	κ: .17, Corr.: .18, p. <.01	**κ: .42, Corr.: .46, p. <.01**	κ: .28, Corr.: .31, p. <.01
body parts	κ: .10 , Corr.: .12 ,p. <.01	κ: .29, Corr.: .32 , p. <.01	**κ: .317, Corr.: .35, p. <.01**	κ: .25, Corr.: .28 , p. <.01
clothes	κ: .07, Corr.: .08, p. <.01	κ: .04, Corr.: .04, n.s.	**κ: .22, Corr.: .26, p. <.01**	κ: .12, Corr.: .14, p. <.01
countries	κ: .10, Corr.: .10, p. <.01	**κ: .43, Corr.: .43, p. <.01**	κ: .39, Corr.: .39, p. <.01	κ: .34, Corr.: .34, p. <.01
fabrics	κ: .11, Corr.: .12, p. <.01	κ: –.08, Corr.: –.09, p. <.05	**κ: .10, Corr.: .12, p. <.01**	κ: .02, Corr.: .02, n.s.
groceries	κ: .13, Corr.: .17, p. <.01	κ: .23, Corr.: .29, p. <.01	**κ: .26, Corr.: .33, p. <.01**	κ: .21, Corr.: .27, p. <.01
hobbies	κ: .15, Corr.: .19, p. <.01	κ: .24, Corr.: .29, p. <.01	**κ: .39, Corr.: .48, p. <.01**	κ: .22, Corr.: .27, p. <.01
insects	κ: .13, Corr.: .20, p. <.01	κ: .08, Corr.: .13, p. <.01	**κ: .15, Corr.: .25, p. <.01**	κ: .11, Corr.: .18, p. <.01
occupations	κ: –.01, Corr.: –.02, n.s.	κ: .09, Corr.: .13, p. <.01	**κ: .17, Corr.: .25, p. <.01**	κ: .15, Corr.: .22, p. <.01
vessels	κ: –.06, Corr.: –.09, p. <.05	κ: .02, Corr.: .02, n.s.	**κ: .12, Corr.: .16, p. <.01**	κ: .08, Corr.: .10, p. <.01

Table 4: Interrater agreement between human annotations and Median Thresholds for each embedding

Categories	Human-Annotated Data versus Embeddings' Median Threshold			
	BERT	**GloVe**	**ConceptNet**	**fastText**
animals	κ: .10, Corr.: .10, p. <.01	κ: .311, Corr.: .31, p. <.01	**κ: .53, Corr.: .54, p. <.01**	κ: .40, Corr.: .40, p. <.01
body parts	κ: –.04 , Corr.: –.037 , n.s.	**κ: .41, Corr.: .41, p. <.01**	κ: .24, Corr.: .24, p. <.01	κ: .38, Corr.: .38, p. <.01
clothes	κ: .07, Corr.: .07, p. <.01	κ: .17, Corr.: .17, p. <.01	**κ: .33, Corr.: .33, p. <.01**	κ: .23, Corr.: .23, p. <.01
countries	κ: .144, Corr.: .164, p. <.01	**κ: .35, Corr.: .39, p. <.01**	κ: .31, Corr.: .36, p. <.01	κ: .30, Corr.: .34, p. <.01
fabrics	**κ: .20, Corr.: .21, p. <.01**	κ: –.08, Corr.: –.08, n.s.	κ: .08, Corr.: .08, n.s.	κ: .10, Corr.: .10, p. <.05
groceries	κ: .17, Corr.: .18, p. <.01	κ: .34, Corr.: .34, p. <.01	κ: .33, Corr.: .34, p. <.01	**κ: .40, Corr.: .41, p. <.01**
hobbies	κ: .17, Corr.: .17, p. <.01	κ: .37, Corr.: .37, p. <.01	**κ: .61, Corr.: .61, p. <.01**	κ: .39, Corr.: .39, p. <.01
insects	κ: .24, Corr.: .27, p. <.01	κ: .23, Corr.: .26, p. <.01	**κ: .40, Corr.: .45, p. <.01**	κ: .29, Corr.: .33, p. <.01
occupations	κ: .03, Corr.: .04, n.s.	κ: .21, Corr.: .23, p. <.01	**κ: .38, Corr.: .40, p. <.01**	κ: .35, Corr.: .37, p. <.01
vessels	κ: –.03, Corr.: –.03, n.s.	κ: .03, Corr.: .03, n.s.	κ: .13, Corr.: .13, p. <.01	**κ: .15, Corr.: .15, p. <.01**

Table 5: (left) GermaNet PB scores, (middle) IC-based relatedness scores, (right) correlation scores between threshold-based sequences and human annotations

Categories	Shortest Path Distance (Mean)	IC-Relatedness Score (Mean)	IC-based Relatedness (with median threshold)
animals	7.11	.27	Corr: –.27, **p.<.01**
body parts	5.99	.21	Corr: –.23, **p.<.01**
clothes	4.08	.31	Corr: –.06, **p.<.01**
countries	2.48	.23	Corr: –.03, n.s.
fabrics	5.48	.30	Corr: –.12, **p.<.05**
groceries	5.38	.26	Corr: –.05, **p.<.01**
hobbies	7.07	.11	Corr: –.44, **p.<.01**
insects	6.37	.37	Corr: –.23, **p.<.01**
occupations	7.61	.20	Corr: –.24, **p.<.01**
vessels	3.46	.26	Corr: –.21, **p.<.01**

ConceptNet and fastText scores with or without GermaNet converge on similar values in the global category.

4.5 Discussion

The results show that automatic prediction of switches aligns best with human annotations when using (i) ConceptNet , (ii) a median-based threshold switch detection and (iii) category-specific thresholds. Strong correlations have been achieved for the *hobbies*, and *animals*, and moderate correlations for the *occupations*, *insects*, *groceries*, *clothes*, *countries* and *body parts* (revisiting Table 4). Performance of automatic switch prediction is worst on *vessels* and *fabrics*. This aligns with the fact that these are the categories resulted with the lowest word counts (see Table 1). The exception here is *insects* with few item produced for this category and subcategory boundaries well represented in the embedding methods. We speculate that this might be due to *insects* occurring in more defined and narrow contexts in the training data, whereas fabrics and vessels may occur in a wider range of contexts. Furthermore, we obtain stronger alignment between automatic prediction and human annotation when taxonomic relations are included via implicit co-learning (e.g. ConceptNet) or explicit integration (e.g. joint metric in Section 4.4). Thus, for this task, taxonomic relations are indispensable and should be part of the automatic scoring mechanisms for better alignment with the human annotations. The selection of threshold methods for defining switch boundaries also plays an important role for getting closer to human annotations. 25-percentile is statistically correlated for almost all categories and representation methods, but overall at weak levels. Decisions based on Median thresholds display stronger alignment, and increase the correlations of all representation methods at a

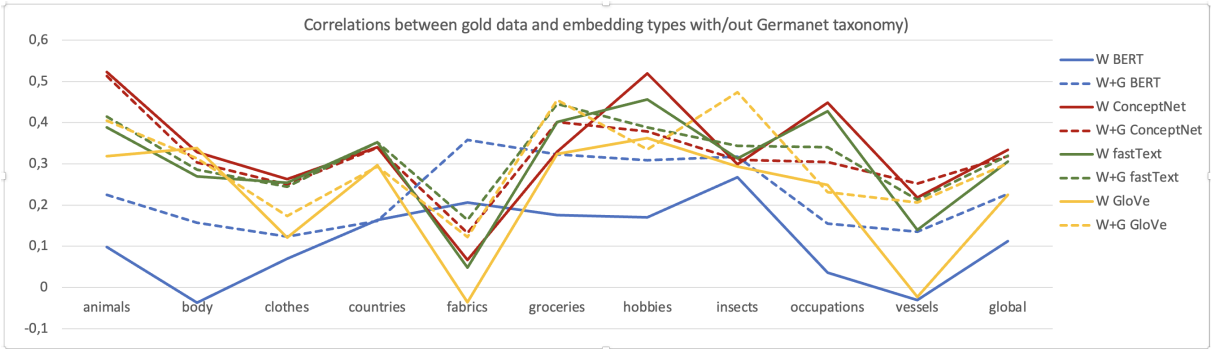

Figure 3: Correlations between the gold annotation and each embedding condition (with/out GermaNet taxonomy.

substantial degree.

The poor results using BERT embeddings could originate from the discrepancy of the task to language modeling, or the lack of training during the conversion from contextualized to static word embeddings. The performance might be improved by increasing the sample size. To train models this large for obtaining static word embeddings seems unreasonably expensive, both computationally and ecologically – especially considering the good performance of simpler approaches enriched with taxonomic relations.

5 Clustering Analysis

The pairwise switch analysis revealed considerable differences between categories as well as favourable results for ConceptNet as a semantic representation. In the following, we report results from an additional clustering-based analysis, in which we take a more global perspective: Instead of investigating adjacent items in fluency sequences, we look at the global semantic organization of all lexical items in the respective categories.

For each category c, we encode the assigned subcategories for individual items into sparse binary vectors. This transformation is necessary for K-Means clustering. This gives us a feature vector with dimensions $(n_{words_c}, n_{subcats_c})$ for human annotations, where n_{words_c} is the number of lexical items and $n_{subcats_c}$ the number of subcategories in the respective category. As simplified, lets assume that we have only three subcategories. A word is represented by the vector [1,0,0] if the word has been assigned to subcategory 1 but not 2 and 3.

We then retrieve the same lexical items from the semantic representations described in Section 3.3, and use each of the feature vectors to fit a *K-Means* clustering model. The k parameter is set depending on the number of annotated subcategories, determined as $k_c = \frac{n_{subcats_c}}{2}$. This parameter reflects the complexity of the domain, i.e., it should allow a large number of clusters for categories with many subcategories. However, there are a few highly fine-grained categories (e.g., fabrics with 28 subcategories for 148 unique words). Therefore, we scaled the value of k for all domains by a constant value.

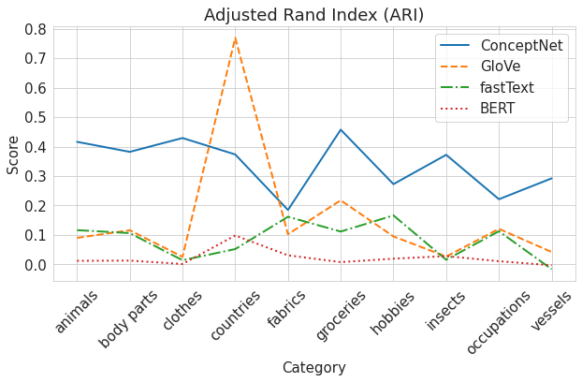

Figure 4: Adjusted Rand Index (ARI) scores for different word embeddings & categories in the fluency data

For evaluation, we rely on the *Adjusted Rand Index (ARI)* (Hubert and Arabie, 1985; Steinley, 2004) for comparing the clusterings based on ConceptNet, GloVe, fasttext and BERT with the results for annotated subcategories. We use the *scikit-learn*[13] library for both clustering and evaluation.

The scores reported in Table 6 and visualized in Figure 4 confirm our previous findings: We see large differences between the categories investigated, with ConceptNet outperforming other semantic representations. One noticeable exception from this is the category *countries*, where GloVe performs surprisingly well. As a possible explanation, we suggest that for different countries the

[13] https://scikit-learn.org/. For the detailed definition of ARI, please visit https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

	ConceptNet	GloVe	fastText	BERT
animals	0.42	0.09	0.12	0.01
body parts	0.38	0.12	0.11	0.01
clothes	0.43	0.03	0.01	0.00
countries	0.37	0.77	0.05	0.10
fabrics	0.18	0.10	0.16	0.03
groceries	0.46	0.22	0.11	0.01
hobbies	0.27	0.10	0.17	0.02
insects	0.37	0.03	0.02	0.03
occupations	0.22	0.12	0.11	0.01
vessels	0.29	0.04	-0.02	0.00

Table 6: Adjusted Rand Index (ARI) scores for different word embeddings & categories in the fluency data

textual context might be very informative. As it is based on lexical co-occurrence, this might result in GloVe representations fairly consistent with human categorization.

6 Conclusion

In this paper, we have explored a range of semantic spaces and switch detection methods for the analysis of the verbal fluency data. To the best of our knowledge, this is the first study that (i) incorporates the taxonomic relations using NLP techniques, (ii) explores a wide variety of semantic categories (10 categories), and (iii) explores the fitness of semantic representations in German for this task.

NLP solutions so far are limited to typical/frequent categories like fruits and animals, leaving the annotation of other categories to laborious manual methods. To develop an automatic scoring mechanism, in-depth analysis for less frequent categories is necessary. Our results revealed various category-specific characteristics.

We showed that choosing individual threshold strategies to detect switch boundaries is essential and a "one-fits-all" solution (using a global threshold) results in less aligned sequences. Still, it can be kept as an option since the degradation is not large for the easy to enumerate categories.

In addition to providing an another perspective to analyze the verbal fluency data for psycholinguistic research, this study also prepares the ground for investigating interesting NLP tasks, like subcategory prediction/generation of the upcoming items. From that perspective, Nighojkar et al. (2022) claims that transformer-based language models perform better on cognitive modeling (more specifically, on predicting the next items given a sequence) than the static approaches. However, their evaluation does not include a comparison to knowledge-enriched models. Although their task differs from exploring semantic spaces to detect category switches, category-specific variations are observed from their results regarding 5 categories (fruits, vegetables, animals, supermarket items, tools, and foods).

Furthermore, applying mere taxonomic relations using a synset taxonomy falls behind embedding methods but proved to be instrumental as an accompanying information source, especially for the hard to enumerate categories.

These results highlight that the task is more challenging than it seems, and we need to go beyond out-of-the-box NLP approaches by understanding the nature of these categories and the task. Future studies aim to improve the integrated switch detection method around the taxonomy-enriched representations using additional modalities and knowledge-graph enriched BERT models.

References

Steffen Aschenbrenner, Oliver Tucha, and Klaus W Lange. 2000. *Regensburger Wortflüssigkeits-Test: RWT*. Hogrefe, Verlag für Psychologie.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 1–10.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Barbara Benigni, Monica Dallabona, Elena Bravi, Stefano Merler, and Manlio De Domenico. 2021. Navigating concepts in the human mind unravels the latent geometry of its semantic space. *Complexity*, 2021.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):1–34.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Christiane Fellbaum. 2010. WordNet. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2018. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 181–187, New Orleans, Louisiana. Association for Computational Linguistics.

Iryna Gurevych and Hendrik Niederlich. 2005. Computing semantic relatedness of germanet concepts. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Proceedings of Workshop" Applications of GermaNet II" at GLDV*, pages 462–474.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Najoung Kim, Jung-Ho Kim, Maria K. Wolters, Sarah E. MacPherson, and Jong C. Park. 2019. Automatic scoring of semantic fluency. *Frontiers in Psychology*, 10.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra Konig. 2017. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, pages 1–7, Montpellier, France).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, volume 26, pages 3111–3119.

Animesh Nighojkar, Anna Khlyzova, and John Licato. 2022. Cognitive modeling of semantic fluency using transformers. In *Proceedings of the 31th International Joint Conference on Artificial Intelligence: Cognitive Aspects of Knowledge Representation*, Vienna, Austria.

Serguei VS Pakhomov and Laura S Hemmy. 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.

Felipe Paula, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. Similarity measures for the detection of clinical conditions with verbal fluency tasks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 231–235, New Orleans, Louisiana. Association for Computational Linguistics.

Luca Pauselli, Brooke Halpern, Sean D Cleary, Benson S Ku, Michael A Covington, and Michael T Compton. 2018. Computational linguistic analysis

applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry research*, 263:74–79.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Mengyang Qiu and Brendan Johns. 2021. A distributional and sensorimotor analysis of noun and verb fluency. *PsyArXiv*.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130.

Zeshu Shao, Esther Janse, Karina Visser, and Antje S. Meyer. 2014. What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5:1–10.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Douglas Steinley. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *neuropsychology*, 11(1):138.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

7 Appendix

7.1 Data Collection Details

- The experiment was conducted online using Qualtrics.
- The task is explained in writing with an accompanying example during the instruction. During the test, they are asked to type words given the category. Typos were corrected where necessary.
- Each of the 24 categories was presented on a separate page. The timer (60sec) started immediately upon presentation. The order is randomized for each participant. Participants needed to click through all pages. Empty results for animals, hobbies, and groceries are not expected. As sanity check, no response for these categories is evaluated as failure (10 participants). Two participants were dropped based on the typing-speed-test. Two participants made free association (e.g., switching from mice to cheese) instead of enumerating within the class.
- Annotators are instructed to follow a commonsense approach: e.g., "cow" would get assigned to a subcategory like "farm animal" but not to its biological taxonomy.
- The annotation process was conducted in two levels. First, the entire world list, which was produced in the experiment, was checked. An annotator could assign a word to multiple categories, e.g., "lion" as "cat-like", "savanna/desert" and "zodiac sign"; these ratings were done without the annotators knowing the context in which the word was produced. In the second run, subcategories for words are checked in the context of a participant's concrete answers, e.g., "lion" in the context of "cat" and "panther" would be assigned to "cat-like" but "lion" in the context of "Capricorn" would be assigned to "zodiac sign."
- Participants were mainly students of psychology receiving credit points
- The study was approved by the ethics board of the Universität Bielefeld.

7.2 Categories

As mentioned in 3.1, based on qualitative and descriptive analysis, we narrow the 24 categories down to those (i) that have at least 75 words produced in total, (ii) which are available in all embeddings' vocabulary list, and (iii) with a minimum average of 5 words per annotated subcategory. It should be noted that categories like *first names* or *computer games* are problematic for distributional

semantic methods. This leaves us with 10 categories marked in bold below.

- "Amphibians": "Amphibian",
- **"Animals": "Tiere",**
- "Body parts": "Körperteile",
- **"Clothes": "Kleidungsstücke",**
- **"Countries": "Länder",**
- "Currencies": "Währungen",
- "Dances": "Tänze",
- **"Fabrics": "Stoffe",**
- "First names": "Vornamen',
- "Flowers": "Blumen',
- "Gods of antiquity": "Götter der Antike",
- **"Groceries": "Lebensmittel",**
- **"Hobbies": "Hobbies",**
- **"Insects": "Insekten",**
- "Metals": "Metalle",
- "Mountains": "Berge",
- **"Occupations": "Berufe",**
- "Precious stones": "Edelsteine",
- "Spices": "Gewürze",
- "Trees": "Baume",
- "Tropical fruits": "tropische Früchte",
- **"Vessels": "Behälter",**
- "Weapons": "Waffen",
- "Wines": "Weinsorten"

Table 7: Subcategory Switch Thresholds for GermaNet IC-Path Based Similarity Scores

	Median	25 Percentile
animals	.28	.25
body parts	.22	.05
clothes	.31	.30
countries	.23	.23
fabrics	.30	.17
groceries	.26	.05
hobbies		
insects	.37	.36
occupations	.20	.12
vessels	.26	.26
global	.25	.12

7.3 Subcategory Switch Thresholds

Table 9 presents the subcategory switch thresholds calculated with respect to median and 25-Percentile values for 4 different embedding spaces.

7.4 Mean Cosine Distance Scores

Average cosine distance scores between pairs across categories and approaches are presented in Table 8.

7.5 GermaNet Distance and Similarity Metrics

Shortest path distance (SD) given category: The shortest path calculation starts with finding the most similar synset for each word in the pair given the category. For example, for the pair <monkey, dog> in the *animal* category, first, the most relevant synsets for the word "monkey" and "dog" for the category *animal* are calculated separately. Later, the minimum path distance between these two synsets is measured.

Path-based (PB) relatedness: Unlike the previous metric that returns absolute path distance between synsets, path-based relatedness measures compute the semantic relatedness between two concepts based on the shortest path between two synsets in the hypernym relation. The shortest path length is the minimal number of nodes forming a path between the two synsets in the relation. It is also useful to disambiguate word senses (e.g. mouse as animal or electronic equipment")

IC-based relatedness. This measure is explained in the main paper (Section).

To illustrate, Figure 5 shows one example sequence produced in the *animal* category with

Table 8: Average cosine distance scores between pairs across categories and approaches. (The numbers in bold format indicates highest similarity within the category, while the underscore indicates second highest scores.

	GloVe	**fastText**	**ConceptNet**	**BERT**
animals	0,29	0,48	0,38	**0,72**
body parts	0,42	0,52	0,44	**0,68**
clothes	0,22	0,45	0,50	**0,72**
countries	0,54	0,56	0,41	**0,70**
fabrics	0,27	0,48	0,45	**0,73**
groceries	0,31	0,48	0,43	**0,70**
hobbies	0,32	0,39	0,26	**0,70**
insects	0,18	0,46	0,49	**0,80**
occupations	0,31	0,43	0,30	**0,68**
vessels	0,23	0,45	0,46	**0,70**

Table 9: Subcategory Switch Thresholds for Word Embeddings

	Median				25 Percentile			
	GloVe	**fastText**	**ConceptNet**	**BERT**	**GloVe**	**fastText**	**ConceptNet**	**BERT**
animals	.28	.48	.37	.71	.17	.38	.21	.64
body parts	.32	.54	.45	.67	.46	.44	.31	.62
clothes	.20	.47	.55	.73	.09	.33	.36	.62
countries	.55	.57	.37	.70	.46	.48	.26	.65
fabrics	.24	.46	.44	.74	.13	.37	.28	.66
groceries	.30	.49	.43	.70	.17	.37	.30	.61
hobbies	.34	.41	.22	.68	.20	.27	.08	.62
insects	.16	.47	.47	0.82	.08	.37	.32	.80
occupations	.35	.46	.29	.67	.19	.35	.16	.61
vessels	.24	.46	.45	.68	.14	.38	.32	.62
global	.31	.49	.41	.71	.17	.37	.24	.63

these above-mentioned GermaNet scores. The first method has no normalization, and although it does a reasonable job for the overall sequence, it returns the same value for <cat, dog> and <rat, mouse> pairs. PB metric addresses the normalization issue, still treats these pairs in a same way. On the other hand, with the inclusion of word frequency values obtained from a large corpus, it becomes more sensitive for these pairs while flattening the other differences in the less frequent items. Since word enumeration during a verbal fluency task results in rare and participant-dependent word pair formations as well as stereotypical pairs, exploring various metrics is instrumental for understanding the task dynamics and developing a technique for automatic scoring.

w1	**w2**	**w3**	**w4-w11**	**w11**	**w12**	**w13**	**w14**	**w15**
Affe (monkey)	*Katze (cat)*	*Hund (dog)*	…	*Pelikan (pelican)*	*Fisch (fisch)*	*Hai (shark)*	*Ratte (rat)*	*Maus (mouse)*

	w1-w2	***w2-w3***	***w3-w4***	***…***	***w11-w12***	***w12-w13***	***w13-w14***	***w14-w15***	…
SD:	8	**2**	11	…	*9*	*5*	*8*	***2***	…
PB:	0.77	**0.94**	0.68	…	*0.85*	*0.88*	*0.77*	***0.94***	…
IC:	0.11	**0.30**	0.04	…	*0.25*	*0.25*	*0.25*	***0.41***	…

Figure 5: Shortest Distance, Path-based and IC-based similarity scores using GermaNet.

Table 10: GermaNet shortest path distance and similarity scores between consecutive synsets

Categories	**Shortest Path Distance**	**PB-based Similarity**	**IC-based Similarity**
animals	**Corr: .21, p.<.01**	**Corr: -.30, p.<.01**	**Corr: –.27, p.<.01**
body parts	**Corr: .19, p.<.01**	**Corr: –.19, p.<.01**	**Corr: –.23 , p.<.01**
clothes	Corr: –.04, n.s.	Corr: .02, n.s.	**Corr: –.06, p.<.01**
countries	**Corr: .077, p.<.01**	**Corr: –.06, p.<.01**	Corr: –.025, n.s.
fabrics	Corr: –.04, n.s.	Corr: –.08, n.s.	**Corr: –.12, p.<.05**
groceries	Corr: .01, n.s.	Corr: –.05, n.s.	**Corr: –.05, p.<.01**
hobbies	**Corr: .282, p.<.01**	**Corr: –.296, p.<.01**	**Corr: –.442, p.<.01**
insects	Corr: .010, n.s.	**Corr: –.18, p.<.05**	**Corr: –.23, p.<.01**
occupations	Corr: .11, n.s.	**Corr: –.11, p.<.05**	**Corr: –.24, p.<.01**
vessels	Corr: .09, n.s.	Corr: –.11, n.s.	**Corr: –.21, p.<.01**