How Shall a Machine Call a Thing?

How shall a machine call a thing?

Federico Torrielli[0000−0001−8037−8828], Amon Rapp[0000−0003−3855−9961], and Luigi Di Caro[0000−0002−7570−637X]

Dept. of Computer Science, University of Torino, Italy
{federico.torrielli, amon.rapp, luigi.dicaro}@unito.it

Abstract. This paper aims to investigate the feasibility of utilising Large Language Models (LLMs) and Latent Diffusion Models (LDMs) for automatically categorising word basicness and concreteness, i.e. two well-known aspects of language having significant relevance on tasks such as text simplification. To achieve this, we propose two distinct approaches: i) a generative Transformer-based LLM, and ii) a image+text multi-modal pipeline, referred to as *stableKnowledge*, which utilises a LDM to map terms to the image level. The evaluation results indicate that while the LLM approach is particularly well-suited for recognising word basicness, *stableKnowledge* outperforms the former when the task shifts to measuring concreteness.

Keywords: Large Language Models · Latent Diffusion Models · Language Basicness · Language Concreteness · Text Simplification.

1 Introduction

Human communication and reasoning rely on lexemes and linguistic expressions that are arranged in hierarchical structures [30]. In this context, researchers in psycholinguistics have identified the concept of *basic level* of language, which refers to the level of inclusiveness that is most efficient for human cognition, as it strikes a balance between information richness and cognitive economy [8]. Basic terms are usually culturally common, salient, or frequently used. Moreover, a variety of studies have consistently found that concrete concepts are easier to identify, recall [23], and understand [33] than abstract ones, supporting the notion that concreteness enhances linguistic processing [36].

The importance of studying and automatically detecting both *basicness* and *concreteness* aspects of a language has a significant impact on several tasks and applications, both of passive and transformative types such as *i)* text complexity analysis [12], *ii)* Word Sense Disambiguation [13], *iii)* Text Simplification [2], *iv)* Machine Translation [19] and others [7,14]. Moreover, automatic tools and novel lexical resources may impact on the education context and/or support the treatment of disorders such as Dyslexia [39].

In this paper, we propose two different approaches built on top of the recent advancement in Natural Language Understanding (NLU) and Computer Vision (CV) technologies, in the specific context of Second Language Acquisition (SLA).

SLA [9] regards language learners (LLs), i.e., adults with a complete process of linguistic (and cognitive) development dealing with the learning of an additional language. While there exists a significant overlap between the two scenarios, LLs are not learning to name new concepts, but rather to assimilate new terms for something they already know how to lexicalise in a native language.

In the context of text simplification, one of the main applications of basic terms, Second Language Acquisition (SLA) holds significant importance. The necessity for simplified texts primarily stems from second language learners' efforts to assimilate new terms for concepts they can already express in their native language, making text simplification particularly relevant to this group. In contrast, native speakers typically possess a more comprehensive understanding of their language, making text simplification less interesting for them. Focusing on SLA in our proposed approaches is thus essential to address second language learners' needs and enhance text simplification as an effective learning tool.

One approach leverages state-of-the-art technology in natural language understanding (NLU), specifically, Transformer-based models known as Large Language Models (LLMs). The hypothesis is that, given their adeptness at processing textual data, they would excel at distinguishing between basic and advanced terminology. The second method involves a multi-modal pipeline that incorporates both text and image processing. The underlying assumption is that abstract concepts are more difficult to represent visually than concrete concepts [17]. Therefore, by first generating synthetic images for abstract and concrete concepts and then attempting to recreate their textual descriptions (using artificially-generated text), this AI pipeline would likely struggle to reconstruct the abstract concepts.

Given these premises, our contribution is thus four-fold:

1. A novel notion of *basicness* for lexical items, inspired by the existing literature on concreteness [35] and realised through an agreement score over a large-scale annotation involving 10 different annotators;
2. A novel resource of *basic*-vs-*advanced* lexicon for the English language (as direct outcome of the previous contribution), composed of 500 open-domain words which includes and extends current basic word lists;
3. A text-based and a multi-modal text+image approach for automatically capturing basicness and concreteness of words by leveraging current state-of-the-art neural architectures in the fields of NLU and CV;
4. An extensive experimentation that both *i)* validates the quality of the novel resource by means of human judgements and *ii)* demonstrates our hypotheses on the effectiveness of the proposed approaches.

The rest of the paper is organised as follows. Section 2 reports notions and principles related to BL and the state of the art in the context of text- and image automatic generation. Section 3 describes the extraction of basic and advanced concepts and the human-in-the-loop creation of a ground truth. Then, Section 4 details the technological pipeline with the obtained results, while Section and 5 concludes the paper with future research directions. Our work materials and the datasets are available at the following link: https://github.com/federicotorrielli/stableKnowledge.

2 Related Work

The idea of identifying basic terms in a language dates back to Rosch et al. in 1976, followed by a large literature proposing an extensive set of names, principles and examples. Then, after the work by Rosch, many measures and detection strategies for BL have been proposed along the years, continuously summarised in specific surveys over time, e.g. in [10] or in the most recent literature on the topic [5]. At the same time, another historical niche in the literature is represented by the work on concept *concreteness*, originated by [25] and later often linked with the ease of processing concrete words in the human mind [35].

Apart from the conceptualisation of basic level (BL, from now on) and concreteness, a number of computational approaches for their automatic detection have been proposed along the year. For example, [22] proposed a set of 52 rules to identify basic level words, working on different characteristics such as the number of characters, prefixes, minimal frequency in SemCor [21] and others. In [11], the authors started from a new set of 518 lemmas belonging to three categories (hand tool, edible fruits and musical instruments) which have been first labelled as basic or not by three annotators. Then, they utilised lexical, structural and frequency-based features to feed standard classification algorithms such as Support Vector Machines and Decision Trees, obtaining an average Cohen's k score of 0.61 with the annotators. More recently, [6] implemented the Rosch's principle of *cue validity* on similar features as in [22], but also employing Distributional Semantics methods and neural architectures (BART [15]), achieving an overall classification accuracy of 75%. Conversely, fewer efforts have been spent on computational methods for concreteness automatic assessment, often leaning towards the creation of dictionaries [4]. However, the utilisation of Large Language Models (LLMs) and Latent Diffusion Models (LDMs) have not been explored in both tasks so far. In this field, generative models like GPT3 [26], BLOOMZ [32] and OPT [41] are considered state-of-the-art for various Language Modeling [34] tasks. These are models based on the Transformer architecture [37], pre-trained on massive text collections that achieve impressive results when generating text. In addition to LLMs, LDMs are increasingly being used for natural language representation tasks, in addition to their traditional use in image synthesis. Examples of such models include DALL-E 2 [27], Stable Diffusion [29] and Imagen [31].

3 A benchmark for *basic* vs *advanced*

While in literature there is a certain agreement over the existence of a basic lexicon, no unique definition actually exists. On the contrary, several notions, principles and frequently-occurring properties have been reported over time. Apart from the proposals of basic lexicons pioneered by Ogden [24] and the many frequency-based vocabularies available, a significant gap within BL-related studies is represented by their weak link to the conceptual level. Indeed, while it is generally assumed to identify and collect basic level *concepts* or categories, all the reported experiments have been mostly made at the conceptual level or

through vague guidelines involving both the lexical and the conceptual areas. In this contribution, we instead manifestly focus on the lexical level, proposing an experimentation with second language learners to empirically grasp an inventory of basic level terms for the English language. In this section, we detail the components of our first contribution, i.e., the creation of a *basicness*-based ground truth.

3.1 Extraction of seed words

In this study, we propose a Transformer-based pipeline for the extraction of basic and advanced words from text. The pipeline is applied to a corpus of literature sources and raw, noisy data from the Internet to create a dataset of 500 seed words. The performance of the pipeline is evaluated by comparing its results to the judgements of ten human annotators who are second language learners.

The first step of the pipeline involves the creation of a corpus of probable basic words (and associated synsets) by extracting them from literature sources and Internet data. A generative Large Language Model is then employed to filter them through the use of a specific Language Model-based prompt. Next, each term is mapped to its corresponding WordNet synset [20] for a subsequent phase of advanced (i.e., non basic) term extraction. The final annotation dataset, consisting of 500 total lemmas, was then obtained through a last selection process. All these phases are detailed in the following paragraphs.

(a) Basic raw list extraction The extraction of basic words from the sources follows simple but clear rules: they must be nouns, not redundant, and easy to learn for a non-English speaker. This was achieved by selecting nouns from basic English word lists such as Ogden's [24] and from language-learning subreddits on Reddit[1]. Then, we used SemCor [21] to map the previously selected nouns into synsets using frequency disambiguation[2]. The resulting basicness raw list is composed of more then 5000 terms, which has been used to test the proposed approaches.

(b) Basic word selection To select a subset of basic words that could be employed in the manual annotation phase, we employed a state-of-the-art LLM, i.e. *OPT-6.7b* [41]. In particular, we hypothesised that a LLM, trained on textual data, would excel in distinguishing between basic and advanced terminology. Our findings confirmed this hypothesis. Since this is a generative transformer model, it was instructed to give a "*yes/no*" output at the prompt "*Is this a simple, basic and short English word that is used in everyday language?*, followed by standard examples mentioned in the literature [5]. Through prompt engineering, we tested several possible prompts, obtaining almost overlapping results.

(c) Advanced word extraction For the identification of advanced terms, we instead proceeded in accordance with the existing approaches (e.g. [11]), i.e.,

[1] `https://www.reddit.com`

[2] For each synset, we selected the noun from its *lemma names* with the highest frequency in SemCor.

by exploring downwards the WordNet sense hierarchy from the selected word list of point (b). A synset for an advanced term is evaluated using four key factors: *i)* lemma-frequency relative to text occurrence, calculated using SemCor; *ii)* limited path distance from the original (basic) synset, measured using path similarity using *nltk* [3]; *iii)* absence of shared words with the hypernym; and *iv)* the absence of basic synsets within the advanced list. With more details:

- (*i*) **significant SemCor frequency**: we look for a hyponym that is rare, but only to a certain extent. An example could be the difference between "*Granny Smith*" and "*Cox's Orange Pippin*" - both are apples, but one is more commonly used in texts than the other;
- (*ii*) **path distance**: since we are traversing WordNet, we used Path Distance to evaluate the similarity instead of a non-native algorithm. The optimal distance from the original basic concept was calculated to be 0.63 through fine-tuning of the results. This condition is necessary since we seek a worthy similarity distance from the basic concept - a good hyponym for "*apple*" must still be an apple;
- (*iii*) **no sharing words between the synset and the hyponym**. This was done to prevent less interesting advanced terms, e.g. as with "*state*" and its hyponym "*American state*", which is probably not the best advanced counterpart among all its hyponyms;
- (*iv*) **no basic words in the advanced list**: we avoided cases where hyponyms can be lexicalised through basic words, e.g., the hyponym of "*ocean*" is "*deep*" which is also a candidate basic word.

We further considered an alternative advanced word extraction method, by direct employing the LLM prompting strategy (thus by asking *OPT* to extract the advanced words). However, the motivation behind pursuing the approach described at this point (c) stemmed from the previous literature claiming that advanced-level words are more frequently identified as hyponyms of a set of selected basic words. [11].

(d) dataset fine-tuning To prepare the dataset for an annotation scenario, a subset of terms was carefully chosen from the resulting list, with a focus on removing any potentially harmful words. This subset was generated by sorting the seed words according to their SemCor frequency and selecting 500 terms, comprising 250 *OPT-basic* and 250 *OPT-advanced* terms. The resulting set was then shuffled. Our dataset of 500 words then underwent human classification (Section 3.2) to establish a gold standard "super-annotator" and subsequently assessed against the latter's judgement (Section 3.4).

3.2 Setting of the Human-based Annotation

In [11], annotators had been asked to mark basic words extracted from the hyponyms of three WordNet synsets[3]. In this contribution, we tried to reshape the experiment without limiting the semantic coverage of the candidate words.

[3] *hand tool*.01, *edible fruit*.01 and *musical instrument*.01

Regarding the methodology and the annotation process through the web interface, we ensured that the scope, method of annotation, and definition of basic words were clearly outlined on the first page of our annotation platform. To aid the annotators in their task, we provided examples of basic and advanced terms on the second page, organised into categories. These examples were presented without definitions or descriptions to avoid any potential bias. Additionally, we included a video that explained the task and provided examples of high-quality annotations from the literature. On the annotation page, we focused on individual words rather than concepts, synsets, or definitions. This approach allowed annotators to indicate whether a word was hard to evaluate, providing valuable feedback for our research.

3.3 Inter-annotation Agreement

We conducted an annotation task on the resulting dataset, recruiting 10 gender-balanced language-learner annotators. These were chosen for their English level (from B1 to C1 in the *CEFR* spectrum) and with different work and study backgrounds. By focusing on individual words and developing a new set of guidelines (see Section 3.2), we were able to achieve a Cohen's κ of 0.70 (with a Krippendorff's alpha of 0.71). The highest value of κ between each pair of annotators was 0.89, while the lowest was 0.66. To further evaluate the reliability of our annotation task along the entire process, we calculated the annotators agreement with a sliding window of 130 words, obtaining the stable sequential values of $0.6834, 0.6255, 0.6268$ and 0.7879.

The annotation task revealed other interesting insights, e.g., the amount of time spent by annotators evaluating a single word[4] rather than specific words that appeared difficult to evaluate, like compound nouns (e.g., *"vitamin pill"*), words borrowed from other languages(e.g., *"avenue"*) and short words that were abstract or conceptually complex, (e.g., *"kin"*).

One aspect of the annotation is that it demonstrated the existence of a *basicness* scale. For example, only a small subset of the whole word list has been annotated as basic or advanced by all annotators with perfect agreement. We could call these words *most basic* and *most advanced* respectively. On the contrary, we identified a *gray area* of $21+14$ (for basic) and $14+21$ (for advanced) terms for which the panel is split 5 to 5. Unsurprisingly, more than 50% of the lexical items falling in this space were marked as *hard* to classify. In Table 1 we summarised different agreements on the annotation, from the mentioned *gray area* cases in the first row to the *most basic/most advanced* cases at the bottom.

3.4 Benchmark Dataset Evaluation

A key step in evaluating the validity of our proposed method was to assess the agreement between human annotators and the list generated with the *OPT*

[4] The results are depicted in this image, which shows that they spent an average of 1.4s on a single word, with 0.9s spent on *OPT*-basic words and 2.0 on the *OPT*-advanced ones.

Annotators split	n. of basic	n. of adv	total
low agreement			
●●●●● vs ●●●●●	35	35	132
●●●●●● vs ●●●●	35	27	
medium agreement			
●●●●●●● vs ●●●	48	43	203
●●●●●●●● vs ●●	52	60	
high agreement			
●●●●●●●● vs ●	35	**48**	165
●●●●●●●●●●	41	41	

Table 1. Our basic vs. advanced agreement distribution over the ten annotators. To make an example (marked in **bold**), 48 advanced words have been classified with a high agreement of 9-vs-1 annotators split.

method (Section 3.1). To this end, we created a baseline "*super-annotator*" by applying majority voting to the basic/advanced annotations. We then compared this newly *super-annotator* annotation with our original *OPT* list and obtained an agreement of $\kappa = 0.63$, with a Precision/Recall/F scores of 0.82/0.81/0.82. This finding suggests that *(i)* large generative language models can effectively differentiate between basic and advanced terms using simple queries, improving the current state of the art by around six percentage F-score points with respect to [6] for the English language, and *(ii)* the LLM demonstrates a strong alignment with the agreement among humans of $\kappa = 0.7$, as reported in Section 3.3 and that could be considered as an asymptotic maximum limit for our task [38].

4 Multi-modal Text+Image Pipeline

In the previous section we focused on the *basicness* aspect of language, providing *i)* a novel benchmark dataset for future and possibly different research objectives and *ii)* demonstrating the capability of a state-of-the-art LLM in classifying basic language. Our second and parallel intent regards language concreteness, a similar and significantly overlapping aspect that, however, has been often faced separately in the current literature. In fact, on the concreteness aspect, different research efforts already carried to benchmark datasets and graded scores for word lists. One of the most employed consists of 4293 nouns in the MRC Psycholinguistic Database [40], where each noun is accompanied by a concreteness score ranging from 0 to 700. By directly testing the LLM-based *OPT* model on the abstract/concrete classification task, and using the best performing concreteness threshold of 380, we reached a very low Cohen's κ of 0.27. This demonstrates that even the most powerful language models are not capable of capturing the hidden different shape behind language concreteness, as opposed to basicness.

Thus, as a further contribution, we introduce a second method operating at the image level with the goal of exploring a more complete and multi-modal perspective by taking advantage of the latest state-of-the-art image synthesis

models in conjunction with LLMs. To the best of our knowledge, this is the first time that such an approach has been proposed and implemented for this task.

4.1 The multi-modal pipeline

In this section, we introduce the pipeline architecture providing an overview of its various components, leveraging state-of-the-art techniques in Natural Language Processing and Computer Vision to enable accurate classification of a wide range of visual and linguistic data. The pipeline is composed of three parts:

- **Image generation step**: given the lemmas, we produce images[5] using Stable Diffusion, a latent diffusion model (LDM) introduced in [29] (see Section 4.1);
- **Interrogation step**: from the images, we extract definitions using BLIP [16], a unified model for vision-language understanding and generation (see Section 4.1);
- **Evaluation step**: to evaluate the similarity between the description produced by the interrogation step and the lemmas in input we used Sentence Transformers [28] embeddings, enabling the use of similarity measures (see Section 4.1).

(a) Image generation using Stable Diffusion The image generation process uses Stable Diffusion 1.5 [29]. In order to maximise performance while simultaneously reducing our carbon footprint, we *i)* enabled the cuDNN auto-tuner for faster convolution, *ii)* utilised the highly performant DPM Solver [18] for efficient model sampling, and *iii)* employed attention slicing, which allows for the computation to be performed in steps rather than all at once[6]. We then generated 5 images per prompt using 30 inference steps and a guidance scale of 7.5. Furthermore, we employ negative prompts [1], such as "*writing, letters, handwriting, words*" to avoid visual clutter resulting from the model attempting to resemble text in the generated images.

(b) BLIP Interrogator module After generating five images for each lemma, we converted them to RGB format and used them as input for the Interrogator module. This module utilises the BLIP large captioning model [16]. Finally, we generated captions using nucleus sampling and a maximum generation length of 20. The synthetic text generation approach utilised in BLIP has been shown to produce results that are comparable to those generated by humans, as demonstrated in various state-of-the-art experiments [16].

(c) Evaluating embeddings using SBERT The final step involved evaluating the quality of the captions we generated by comparing them to the original lemmas. To accomplish this, we utilised SBERT [28] to produce embeddings where to apply cosine similarity, thus having a quantitative evaluation measure on the generated captions. To further refine our analysis, we experimented with both the mean and the maximum similarity value across all five captions.

[5] In-depth examples of the outputs can be examined in the following link: https://github.com/federicotorrielli/stableKnowledge/tree/master/appendix_b.pdf

[6] The total GPU hours required for image generation, captioning and evaluation was approximately 12 hours using a single consumer grade NVIDIA 2080Ti.

4.2 Language vs Vision technologies: an evaluation of the two approaches on basicness and concreteness

By pairing the results on concreteness classification with the LLM-based approach, the *stableKnowledge* pipeline demonstrated superior performance, achieving a Cohen's κ of **0.57** (more than the double) with a concreteness threshold of 520 and a cosine similarity threshold of 0.3564, as detailed in Table 2. On the contrary, the LLM-based method outperformed the LDM-based pipeline *stableKnowledge* in classifying language basicness, as shown in Table 3. Thus, our initial hypotheses are fully verified by the experiments.

Furthermore, it is worth to outline some relation between the two separate dimensions of basicness and concreteness. By looking at the accuracy values in Tables 2 and 3, *OPT* experiences a significant drop in performance uniquely on the abstract concepts. Contrarily, it performs at best with basic, advanced but also concrete expressions.

	OPT model ($k = 0.27$)			*stableKnowledge* ($k = 0.57$)		
	Precision	Recall	F1	Precision	Recall	F1
abstract	0.63	0.40	0.49	**0.77**	**0.85**	**0.81**
concrete	0.70	**0.85**	**0.77**	**0.82**	0.72	0.76

Table 2. Accuracy scores of *OPT* and *stableKnowledge* on the abstract/concrete task.

	OPT model ($k = 0.63$)			*stableKnowledge* ($k = 0.21$)		
	Precision	Recall	F1	Precision	Recall	F1
basic	**0.82**	**0.81**	**0.82**	0.61	0.59	0.60
advanced	**0.81**	**0.82**	**0.82**	0.59	0.61	0.60

Table 3. Accuracy scores of *OPT* and *stableKnowledge* on the basic/advanced task.

5 Conclusion and Future Work

In this paper, we aimed to build upon the existing literature and techniques related to the classification of the basic nature of a language by developing a novel notion of *basicness*, i.e. a graded representation obtained through human-in-the-loop experiments, taking inspiration from the existing works on *concreteness*. First, we generated a novel candidate word list integrating existing principles and resources, which resulted to overcome the current state of the art in terms of its open-domain and balanced qualities. Then, we proposed a human-in-the-loop methodology for the realisation of the basicness idea through a 10-annotators panel, reaching the highest human agreement scores as compared with the current literature, up to 0.89 of Fleiss' *k*. Finally, we experimented with the current state-of-the-art approaches in Natural Language Understanding and in Computer Vision on the automatic classification of both basicness and concreteness, establishing new standards on the topic and highlighting the power of generative models on the two tasks. Future work includes *i)* applying the presented approaches to other language-oriented tasks, *ii)* examining the psycholinguistic implications of *basicness* and *iii)* experimenting with models hyperparameters.

References

1. Negprompt (2023), https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt
2. Al-Thanyyan, S.S., Azmi, A.M.: Automated text simplification: a survey. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)
3. Bird, S., Loper, E.: NLTK: The natural language toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions. pp. 214–217. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
4. Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods **46**, 904–911 (2014)
5. Castellanos, A., Tremblay, M.C., et al.: Basic classes in conceptual modeling: theory and practical guidelines. Journal of the Association for Information Systems **21**(4), 3 (2020)
6. Chen, Y., Teufel, S.: Synthetic textual features for the large-scale detection of basic-level categories in English and Mandarin. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8294–8305. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)
7. Di Caro, L., Ruggeri, A.: Unveiling middle-level concepts through frequency trajectories and peaks analysis. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 1035–1042 (2019)
8. Finton, D.J.: Cognitive economy and the role of* representation in on-line learning. The University of Wisconsin-Madison (2002)
9. Gass, S.M., Behney, J., et al.: Second language acquisition: An introductory course. Routledge (2020)
10. Hajibayova, L.: Basic-level categories: A review. Journal of Information Science **39**(5), 676–687 (2013)
11. Hollink, et al.: Predicting the basic level in a hierarchy of concepts. In: Research Conference on Metadata and Semantics Research. pp. 22–34. Springer (2020)
12. Jensen, K.T.: Indicators of text complexity. Mees, IM; F. Alves & S. Göpferich (eds.) pp. 61–80 (2009)
13. Lacerra, C., Bevilacqua, M., et al.: Csi: A coarse sense inventory for 85% word sense disambiguation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 8123–8130 (2020)
14. Leone, V., Siragusa, G., Di Caro, L., Navigli, R.: Building semantic grams of human knowledge. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2991–3000 (2020)
15. Lewis, M., Liu, Y., et al.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
16. Li, J., Li, D., et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022)
17. Löhr, G.: What are abstract concepts? on lexical ambiguity and concreteness ratings. Review of Philosophy and Psychology **13**(3), 549–566 (2022)
18. Lu, C., Zhou, Y., et al.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps (2022)
19. Marchisio, K., Guo, J., et al.: Controlling the reading level of machine translation output. In: Proceedings of Machine Translation Summit XVII: Research Track. pp. 193–203 (2019)

20. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
21. Miller, G.A., Chodorow, M., et al.: Using a semantic concordance for sense identification. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994)
22. Mills, C., Bond, F., et al.: Automatic identification of basic-level categories. In: Proceedings of the 9th Global Wordnet Conference. pp. 298–305 (2018)
23. Nelson, D.L., Schreiber, T.A.: Word concreteness and word structure as independent determinants of recall. Journal of memory and language **31**(2), 237–260 (1992)
24. Ogden, C.K.: Basic english: A general introduction with rules and grammar (1930)
25. Paivio, A.: Abstractness, imagery, and meaningfulness in paired-associate learning. Journal of Verbal Learning and Verbal Behavior **4**(1), 32–38 (1965)
26. Radford, A., Kim, J.W., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
27. Ramesh, A., Dhariwal, P., et al.: Hierarchical text-conditional image generation with clip latents. ArXiv **abs/2204.06125** (2022)
28. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019)
29. Rombach, R., Blattmann, A., et al.: High-resolution image synthesis with latent diffusion models (2021)
30. Rosch, E., Mervis, C.B., et al.: Basic objects in natural categories. Cognitive psychology **8**(3), 382–439 (1976)
31. Saharia, C., Chan, W., et al.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022)
32. Scao, T.L., Fan, A., et al.: Bloom: A 176b-parameter open-access multilingual language model. ArXiv **abs/2211.05100** (2022)
33. Schwanenflugel, P.J., Shoben, E.J.: Differential context effects in the comprehension of abstract and concrete verbal materials. Journal of Experimental Psychology: Learning, Memory, and Cognition **9**(1), 82 (1983)
34. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 623–656 (1948)
35. Solovyev, V.: Concreteness/abstractness concept: State of the art. In: Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 10-16, 2020, Moscow, Russia 9. pp. 275–283. Springer (2021)
36. Strain, E., Patterson, K., et al.: Semantic effects in single-word naming. Journal of Experimental Psychology: Learning, Memory, and Cognition **21**(5), 1140 (1995)
37. Vaswani, A., Shazeer, N.M., et al.: Attention is all you need. ArXiv **abs/1706.03762** (2017)
38. Warrens, M.J.: Five ways to look at cohen's kappa. Journal of Psychology & Psychotherapy **5**(4), 1 (2015)
39. Washburn, E.K., Joshi, R.M., et al.: Teacher knowledge of basic language concepts and dyslexia. Dyslexia **17**(2), 165–183 (2011)
40. Wilson, M.: Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. Behavior research methods, instruments, & computers **20**(1), 6–10 (1988)
41. Zhang, S., Roller, S., et al.: Opt: Open pre-trained transformer language models (2022)