Visual groundedness as an organizing principle for word class: Evidence from Japanese

Anonymous ACL submission

1 Introduction

Since classical times, one of the fundamental ideas in linguistic theory is that words are divided into categories with shared syntactic and morphological behaviour. Often called "word classes" or "parts of speech", these classes represent an intersection between linguistic form and semantic function. For example, nouns prototypically refer to objects, and verbs to actions or events.

What is the theoretical status of the relationship between meaning and word class? Within any word class in a given language, exceptions to their semantic properties abound. Nevertheless, there is a great degree of cross-linguistic consistency in the relationship between the meaning of lexical items and their syntactic behaviour–the vast majority of languages clearly handle object words differently from action words. Property words also tend to have special morphosyntactic expression across languages, differing from both nouns and verbs. But for each of these distinctions, there are languages where it is not clearly relevant (Bisang, 2010).

How can a theory explain both these strong universal tendencies and well-established deviations from them? Recent work in computational linguistics has attempted to formalize aspects of the relationship between meaning and form (Rauhut, 2023; Haley et al., 2024). In this work, we focus on Haley et al. (2024)'s notion of (visual) *groundedness*. Groundedness formalizes the notion of how much information a word conveys about an utterance's "meaning" in context–how meaningful vs. grammatical a word is. Haley et al. (2024) showed that visual groundedness shows a clear relationship to the distinction between lexical and functional word classes across 30 languages, demonstrating substantial cross-linguistic consistency–the same classes have similar groundedness across languages. Notably, nouns > adjectives > verbs in terms of visual groundedness, despite all being lexical classes.

If word classes are organized in part by the (visual) groundedness of the meanings they express, then variation in word classes should be associated with differences in groundedness of the expressed meanings. In this study, we focus on Japanese property words, which have the unusual property of constituting two formally very distinct word classes, rather than a single "adjective" class. Building on the insight that one of these classes is more formally "nominal" (*na*-adjectives) and one more "verbal" (*i*-adjectives), we hypothesise that we should see analogous trends in function: one class serving more prototypically nominal functions and one more prototypically verbal. In terms of visual groundedness, this corresponds to higher values for the nominal class.

2 Method

Groundedness is formally defined as the pointwise mutual information between a word/linguistic unit in the context of an utterance, and the meaning of that utterance. We focus on *visual groundedness*–representing meaning with an image. This simplifying assumption makes estimating (visual) groundedness with existing datasets and neural models tractable, and has interesting connections to relevant notions like imageability and perceptual strength. In particular, for an image I and word w_t in an utterance $W = w_1, w_2, w_3 ... w_t ...$, we formalise groundedness as:

$$\text{Groundedness}(w_t) = \log p(w_t \mid I, \mathbf{w}_{<t}) - \log p(w_t \mid \mathbf{w}_{<t}) \quad (1)$$

This allows us to compute groundedness as a *difference in surprisal* between an image captioning model and a (domain-matched) language model.

We use the model released by Haley et al. (2024) as a language model and PaliGemma as the image captioning model. We use the `sudachipy`[1] part

[1]

of speech tagger to tag words as *i*-adjectives and *na*-adjectives. We focus on the Crossmodal-3600 (XM3600) dataset (Thapliyal et al., 2022), because of its high quality of manual captioning.

As noted by Haley et al. (2024), single groundedness estimates can be noisy, so we filter for only adjective types which occur at least 5 times in our corpus. This is especially important as *na*-adjectives are less frequent than *i*-adjectives in our corpus.

3 Results

Across our corpus of 7185 captions, we find 399 *na*-adjective tokens and 3058 *i*-adjective tokens. These tokens belong to 42 *i*-adjective types and 26 *na*-adjective types. On average, the *na*-adjectives display higher groundedness than *i*-adjectives (3.41 vs. 1.98). Our data has a nested structure, with many tokens of a single word type, and this word type influences groundedness independently of word class (*i*-adjective vs. *na*-adjective). To better estimate the effect of word class itself, we use a linear mixed effects model, with fixed effects of position and word class and a random effect for word type. Under this model, we find a significant effect of word class ($p = 0.029$). Specifically, we find that *na*-adjectivehood increases groundedness by 0.89 ± 0.40 bits.

Two terms are used to compute our visual groundedness measure: surprisal under a language model and surprisal under an image captioning model. Is the association between groundedness and the word class distinction above primarily due to one of these terms? Of particular concern is the first term: perhaps *na*-adjectives are just *a priori* more surprising in the linguistic signal (e.g. expressing lower-frequency concepts). If we find a strong correlation between word class and LM surprisal, it may be that the information provided by the image is dominated by these effects. Fitting the same fixed and random effects as before to instead predict LM surprisal, we do not find a significant effect ($p = 0.133, \beta = 1.17 \pm 0.77$). Similarly, we do not find a significant effect of word class on the captioning surprisal alone ($p = 0.591, \beta = 0.38 \pm 0.61$). So it is only through the interaction between these two factors (groundedness) that an association with word class emerges.

4 Conclusion

Together, our results suggest that *na*-adjectives are used to express more visually grounded meanings than *i*-adjectives in Japanese. In contrast to prior work which failed to find a semantic organizing principle for this distinction (Morita, 2010; Oshima et al., 2019), our work suggests that the formal similarities *i*-adjectives and *na*-adjectives display to verbs and nouns respectively are not arbitrary, but reflect their semantic character.

While still exploratory, our results suggest an exciting role for groundedness in computational linguistics. Together with Haley et al. (2024), these results point to the utility of groundedness not just for explaining cross-linguistic *consistency* in word class organization, but also *variation*. Beyond this, groundedness can also be a useful tool for framing and answering questions about the relationship between form and meaning in a particular language, not just cross-linguistically. While groundedness is only somewhat correlated with norms like concreteness or imageability, concreteness allows the asking of related questions where such norms are not available–no relevant concreteness or imageability norms exist for Japanese adjectives. Future work should further validate these results on a larger array of words and datasets, and with new and improved models, and also explore such traditional, human-annotated norms.

References

Walter Bisang. 2010. Word Classes. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

Coleman Haley, Sharon Goldwater, and Edoardo Ponti. 2024. A grounded typology of word classes. *Preprint*, arXiv:2412.10369.

Chigusa Morita. 2010. The internal structures of adjectives in japanese. *Linguistic Research*, 26:105–117.

David Oshima, Kimi Akita, and Shun Sano. 2019. Gradability, scale structure, and the division of labor between nouns and adjectives: The case of japanese. *Glossa: a journal of general linguistics*, 4(1):41.

Alexander Rauhut. 2023. *Quantitative Aspects of the Word Class Continuum in English*. Ph.D. thesis, Freie Universität Berlin.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.