Distributional Interaction of Concreteness and Abstractness in Verb–Noun Subcategorisation

Diego Frassinelli, Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
[frassinelli|schulte]@ims.uni-stuttgart.de

Abstract

In recent years, both cognitive and computational research has provided empirical analyses of contextual co-occurrence of concrete and abstract words, partially resulting in inconsistent pictures. In this work we provide a more fine-grained description of the distributional nature in the corpus-based interaction of verbs and nouns within subcategorisation, by investigating the concreteness of verbs and nouns that are in a specific syntactic relationship with each other, i.e., subject, direct object, and prepositional object. Overall, our experiments show consistent patterns in the distributional representation of subcategorising and subcategorised concrete and abstract words. At the same time, the studies reveal empirical evidence why contextual abstractness represents a valuable indicator for automatic non-literal language identification.

1 Introduction

The need of providing a clear description of the usage of concrete and abstract words in communication is becoming salient both in cognitive science and in computational linguistics. In the cognitive science community, much has been said about concrete concepts, but there is still an open debate about the nature of abstract concepts (Barsalou and Wiemer-Hastings, 2005; McRae and Jones, 2013; Hill et al., 2014; Vigliocco et al., 2014). Computational linguists have recognised the importance of investigating the concreteness of contexts in empirical models, for example for the automatic identification of non-literal language usage (Turney et al., 2011; Köper and Schulte im Walde, 2016; Aedmaa et al., 2018).

Recently, multiple studies have focussed on providing a fine-grained analysis of the nature of concrete vs. abstract words from a corpus-based perspective (Bhaskar et al., 2017; Frassinelli et al., 2017; Naumann et al., 2018). In these studies, the authors have shown a general but consistent pattern: concrete words have a preference to co-occur with other concrete words, while abstract words co-occur more frequently with abstract words. Specifically, Naumann et al. (2018) performed their analyses across parts-of-speech by comparing the behaviour of nouns, verbs and adjectives in large-scale corpora. These results are not fully in line with various theories of cognition which suggest that both concrete and abstract words should co-occur more often with concrete words because concrete information links the real-world usage of both concrete and abstract words to their mental representation (Barsalou, 1999; Pecher et al., 2011).

2 The Current Study

In the current study we build on prior evidence from the literature and perform a more fine-grained corpus-based analysis on the distribution of concrete and abstract words by specifically looking at the types of syntactic relations that connect nouns to verbs in sentences. More specifically, we look at the concreteness of verbs and the corresponding nouns as subjects, direct objects and prepositional objects. This study is carried out in a quantitative fashion to identify general trends. However, we also look into specific examples to better understand the types of nouns that attach to specific verbs.

First of all, we expect to replicate the main results from Naumann et al. (2018): in general, concrete nouns should co-occur more frequently with concrete verbs and abstract nouns with abstract verbs. Moreover, we expect to identify the main patterns that characterise semantic effects of an interaction of concreteness in verb-noun subcategorisation, such as collocations and meaning shifts.

The motivation for this study is twofold: (1) From a cognitive science perspective we seek additional and more fine-grained evidence to better understand the clash between the existing corpus-based studies and the theories of cognition which predict predominantly concrete information in the context of both concrete and abstract words. (2) From a computational perspective we expect some variability in the interaction of concreteness in verb-noun subcategorisation, given that abstract contexts are ubiquitous and salient empirical indicators for non-literal language identification, cf. *carry a bag* vs. *carry a risk*.

3 Materials

In the following analyses, we used nouns and verbs extracted from the Brysbaert et al. (2014) collection of concreteness ratings. In this resource, the concreteness of 40,000 English words was evaluated by human participants on a scale from 1 (abstract) to 5 (concrete).

Given that participants did not have any overt information about part-of-speech (henceforth, POS) while performing the norming study, Brysbaert et al. added this information post-hoc from the SUBTLEX-US, a 51-million word subtitle corpus (Brysbaert and New, 2009). In order to align the POS information to the current study, we disambiguated the POS of the normed words by extracting their most frequent POS from the 10-billion word corpus ENCOW16AX (see below for details). Moreover, as discussed in previous studies by Naumann et al. (2018) and Pollock (2018), mid-range concreteness scores indicate words that are difficult to categorise unambiguously regarding their concreteness. For this reason and in order to obtain a clear picture of the behaviour of concrete vs. abstract words, we selected only words with very high (concrete) or very low (abstract) concreteness scores. We included in our analyses the 1000 most concrete (concreteness range: 4.86 – 5.00) and 1000 most abstract (1.04 – 1.76) nouns, and the 500 most concrete (3.80 – 5.00) and most abstract (1.19 – 2.00) verbs. We chose to include a smaller selection of verbs compared to the nouns because we considered verbs to be more difficult to evaluate by humans according to their concreteness scores and consequently noisier and more ambiguous for the analyses we are conducting.

The corpus analyses were performed on the parsed version of the sentence-shuffled English ENCOW16AX corpus (Schäfer and Bildhauer, 2012). For each sentence in the corpus, we extracted the verbs in combination with the nouns when they both occur in our selection of words from Brysbaert et al. (2014) and when the nouns are parsed as subjects (in active and passive sentences: *nsubj* and *nsubjpass*), direct objects (*dobj*) or prepositional objects (*pobj*) of the verbs. In the case of *pobj*, we considered the 20 most frequent prepositions (e.g., *of, in, for, at*) in the corpus.

In total, we extracted 11,716,189 verb-noun token pairs including 3,814,048 abstract verb tokens; 7,902,141 concrete verb tokens; 3,701,669 abstract noun tokens; and 8,014,520 concrete noun tokens. In 2,958,308 cases, the noun was parsed as the subject of the verb (with 748,438 of them as subjects in passive constructions), in 5,011,347 cases the noun was the direct object, and in 3,746,534 cases the noun was a prepositional object. Already by looking at these numbers it is possible to identify a strong frequency bias in favour of concrete words; we will discuss later in the paper how this bias affects the results reported. All the analyses reported in the following sections are performed at token level.

4 Quantitative Analysis

In a pre-test we analysed the overall distributions of verbs and nouns according to their concreteness scores. Figure 1 shows the overall distributions of verbs (left, M=3.4, SD=1.1) and nouns (right, M=3.9, SD=1.6) included in our analyses. Overall, nouns have significantly more extreme values than verbs: the majority of concrete nouns have concreteness scores clustering around 5.00 while concrete verbs cluster around 4.0. Similarly, abstract nouns have significantly lower scores (i.e., they are more abstract) than

Function	Abstract Verbs	Concrete Verbs	Difference C-A	Overall
nsubj	3.57 (± 1.65)	4.41 (± 1.22)	0.84***	4.07 (± 1.46)
nsubjpass	3.34 (± 1.68)	4.20 (± 1.39)	0.86***	3.85 (± 1.56)
dobj	2.65 (± 1.58)	4.30 (± 1.31)	1.65***	3.76 (± 1.60)
pobj	3.10 (± 1.66)	4.20 (± 1.38)	1.10***	3.91 (± 1.54)
in	3.06 (± 1.65)	4.37 (± 1.25)	1.31***	4.01 (± 1.49)
at	2.58 (± 1.51)	4.11 (± 1.24)	1.53***	3.79 (± 1.58)
for	2.86 (± 1.64)	3.36 (± 1.69)	0.50***	3.15 (± 1.69)
of	3.21 (± 1.67)	4.23 (± 1.36)	1.02***	3.92 (± 1.53)

Table 1: Mean concreteness scores (± standard deviation) and differences between the nouns subcategorised by concrete vs. abstract verbs within a specific syntactic function.

abstract verbs. The numerical difference in the presence of extreme scores is also highlighted by the much higher standard deviation characterising nouns compared to verbs. We interpret the lower amount of "real" extremes (1 and 5) for verbs as an indicator of the difficulty that participants had to clearly norm verbs compared to nouns. For example, while comparing the nouns $belief_{1.2}$ and $ball_{5.0}$ humans would have a clear agreement on highly abstract and highly concrete scores; on the contrary, the distinction between $moralise_{1.4}$ and $sit_{4.8}$ might be less clear.[1]

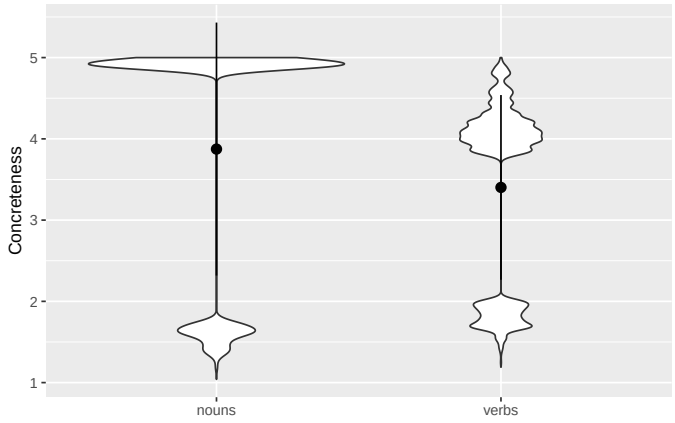
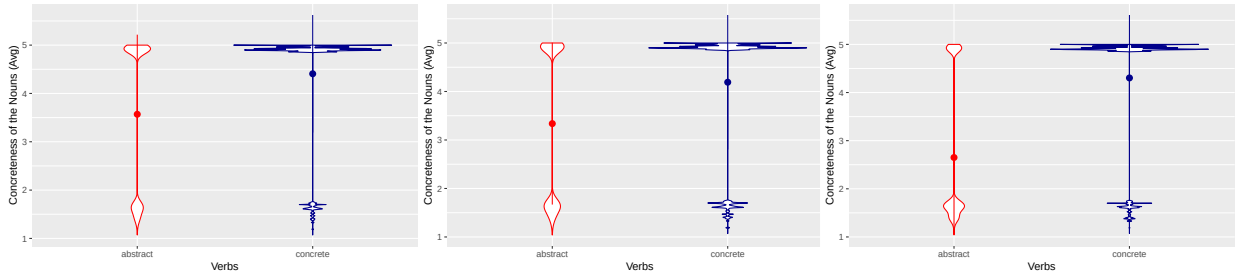


Figure 1: Overall distribution of concreteness scores for nouns (left) and verbs (right). The dots indicate the mean values and the solid vertical lines mark the standard deviations.

In our main study, we analysed the concreteness of the nouns that are in a specific and direct syntactic relation with verbs. The overall distributions in Figure 2 are extremely consistent across syntactic relations: when looking at the means, the concreteness of nouns subcategorised by concrete verbs is significantly higher than the concreteness of nouns subcategorised by abstract verbs (all p-values < 0.001). This result is perfectly in line with the more general analysis by Naumann et al. (2018).

Table 1 investigates more deeply the interaction between the concreteness of verbs and nouns for different syntactic functions. It reports the average concreteness scores of the nouns subcategorised by concrete and abstract verbs (± standard deviation), the difference between the concrete and abstract scores (with significance tests) and the overall average concreteness score by function. The statistical analyses have been performed using a standard linear regression model. The comparison between the scores in the first two columns (Abstract Verbs and Concrete Verbs) confirms that subject and direct object nouns that are subcategorised by concrete verbs are significantly more concrete than those subcategorised by abstract verbs. The "Difference C-A" column shows that these differences are all highly significant. In addition, the nouns subcategorised by concrete verbs are extremely high on the concreteness scale (mean

[1] In this paper the number in subscript indicates the concreteness score from the Brysbaert et al. (2014) norms.



(a) Active subject nouns. (b) Passive subject nouns. (c) Direct object nouns.

Figure 2: Distribution of concreteness scores for the nouns subcategorised by abstract (left/red) and concrete (right/blue) verbs in different syntactic functions. The dots indicate the mean values and the solid vertical lines mark the standard deviations.

values: 4.2 – 4.41) while the nouns subcategorised by abstract verbs have only mid-scores (mean values: 2.65 – 3.57).

By zooming in on the specific functions, we see that subjects are significantly more concrete than direct objects for both abstract and concrete verbs. The concreteness scores of subjects of passivised sentences are in between in both categories. This pattern is confirmed by looking at the "Overall" column.

Prepositional objects that are subcategorised by concrete verbs are significantly more concrete than prepositional objects subcategorised by abstract verbs, across prepositions. However, given the extreme variability in the prepositions used, we will analyse the most representative *pobjs* more specifically in the following section.

5 Qualitative Analysis

In order to better understand the patterns of concreteness behind each syntactic function introduced in the previous section, we performed a series of qualitative analyses, by looking at the most frequent verb-noun combinations grouped by syntactic function. For both functions *nsubj* and *dobj* we see the same strong pattern as in the general analyses in Section 4: concrete verbs have a strong overall preference for concrete complements (*map*$_{4.9}$ *show*$_{4.0}$, *boil*$_{4.2}$ *water*$_{5.0}$). Regarding abstract verbs, we find a preference for subcategorising abstract direct objects (*reduce*$_{2.0}$ *risk*$_{1.6}$), but -in contrast- a preference for concrete subjects (*student*$_{4.9}$ *need*$_{1.7}$). Appropriately, surface subjects in passivised clauses have preferences that are in between those for surface subjects and direct objects in active clauses, presumably because they are semantically comparable to the direct objects of the action encoded by the corresponding verb.

When looking into exceptions to this predominant pattern, we find collocations and non-literal language, such as metaphors and metonyms. For example, metaphorical language usage occurs when concrete verbs attach to abstract direct objects (*carry*$_{4.0}$ *risk*$_{1.6}$ vs. *carry*$_{4.0}$ *bag*$_{4.9}$, *catch*$_{4.1}$ *moment*$_{1.6}$ vs. *catch*$_{4.1}$ *insect*$_{4.9}$); while abstract verbs collocated with concrete direct objects trigger a metonymical use (*recommend*$_{1.7}$ *book*$_{4.9}$ vs. *write*$_{4.2}$ *book*$_{4.9}$).

When looking at prepositional objects it is possible to identify three main behaviours: i) a main preference for concrete verbs and nouns (e.g., "in" and "at"); ii) a strong interaction with abstract verbs and nouns (e.g., "for"); iii) a mixed co-occurrence with both concrete and abstract verbs and nouns (e.g.,"of"). The following paragraphs report a qualitative discussion about the predominant verbs and nouns with regard to the four prepositions "in", "at", "for", and "of".

The preposition *in* manifests a very strong interaction with concrete verbs and concrete nouns. Some examples among the most frequent ones in the corpus are: *write*$_{4.2}$ *in book*$_{4.9}$ and *sleep*$_{4.4}$ *in bed*$_{5.0}$. The only rare exceptions to this pattern refer to idiomatic structures like: *carry*$_{4.0}$ *in accordance*$_{1.5}$ or *carry*$_{4.0}$ *in manner*$_{1.6}$. Table 1 confirms that the preposition *in* triggers very high concreteness scores in general and the highest concreteness scores for nouns that are subcategorised by concrete verbs.

The preposition *at* connects mainly concrete verbs with concrete nouns: *sit*$_{4.8}$ *at table*$_{4.9}$ and *eat*$_{4.4}$ *at restaurant*$_{4.9}$. However, in strong collocations it shows a preference for abstract nouns: *jump*$_{4.5}$ *at chance*$_{1.6}$ or *happen*$_{1.8}$ *at moment*$_{1.6}$. This pattern is confirmed by Table 1 too, where concrete verbs have high scores while abstract verbs have the lowest scores in the entire table.

The preposition *for*, on the other hand, mainly occurs with abstract nouns that are subcategorised by abstract verbs: *need*$_{1.7}$ *for purpose*$_{1.5}$ and *imagine*$_{1.5}$ *for moment*$_{1.6}$. Exceptions to this pattern are due to metonymic readings like *write*$_{4.2}$ *for magazine*$_{5.0}$ and *run*$_{4.3}$ *for office*$_{4.9}$. Correspondingly, we see the lowest overall concreteness score across verbs in Table 1.

Finally, the preposition *of* shows a mixed interaction in the concreteness of verbs and nouns. This preposition co-occurs mainly with very concrete verbs that however subcategorise both highly concrete nouns (*run*$_{4.3}$ *of water*$_{5}$) but also highly abstract nouns (*run*$_{4.3}$ *of idea*$_{1.6}$) in cases of metaphorical use. As expected, the overall concreteness for this function in Table 1 is among the highest both for concrete and abstract verbs.

6 General Discussion & Conclusion

The aim of this study was to provide a fine-grained empirical analysis of the concreteness nature in verb-noun subcategorisation. The general pattern already described in Naumann et al. (2018) is confirmed by our quantitative analysis: overall, concrete verbs predominantly subcategorise concrete nouns as subjects and direct objects, while abstract verbs predominantly subcategorise abstract nouns as subjects and direct objects. A qualitative analysis revealed that exceptions to the predominant same-class interaction indicate semantic effects in verb-noun interaction: collocation, metaphor and metonymy, which shows the usefulness of detecting abstractness in the contexts of verbs as salient features in automatic non-literal language identification.

A slightly more variable pattern emerges when looking at prepositional objects. We identified three main clusters of prepositions that behave differently according to their preferred nouns and verbs. The prepositions in the first cluster (e.g., "in" and "at") co-occur mostly with concrete verbs and nouns; the prepositions in the second cluster (e.g., "for") have a strong preference for abstract verbs and nouns; while the prepositions in the third cluster (e.g, "of") show variability in the concreteness of the related nouns. Once again, the divergence form the general pattern is often ascribable to cases of non-literal language.

This study, on the one hand, provided additional and more fine-grained evidence of the clash between the existing corpus-based studies and the theories of cognition which predict predominantly concrete information in the context of both concrete and abstract words. This was achieved by zooming in on the contexts which stand in a direct syntactic relation to the target word. In addition, they provided useful indicators to the implementation of computational models for the automatic identification and classification of non-literal language.

References

Aedmaa, E., M. Köper, and S. Schulte im Walde (2018). Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-Literal Usage of Estonian Particle Verbs. In *Proceedings of the NAACL 2018 Student Research Workshop*, New Orleans, LA, USA, pp. 9–16.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences 22*, 577–660.

Barsalou, L. W. and K. Wiemer-Hastings (2005). Situating Abstract Concepts. In D. Pecher and R. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, Chapter 7, pp. 129–163. New York: Cambridge University Press.

Bhaskar, S. A., M. Köper, S. Schulte im Walde, and D. Frassinelli (2017). Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns. In *Proceedings of the IWCS Workshop on Foundations of Situated and Multimodal Communication*, Montpellier, France.

Brysbaert, M. and B. New (2009). Moving beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods 41*(4), 977–990.

Brysbaert, M., A. B. Warriner, and V. Kuperman (2014). Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods 64*, 904–911.

Frassinelli, D., D. Naumann, J. Utt, and S. Schulte im Walde (2017). Contextual Characteristics of Concrete and Abstract Words. In *Proceedings of the 12th International Conference on Computational Semantics*, Montpellier, France.

Hill, F., A. Korhonen, and C. Bentz (2014). A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science 38*(1), 162–177.

Köper, M. and S. Schulte im Walde (2016). Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA, pp. 353–362.

McRae, K. and M. Jones (2013). Semantic Memory. *The Oxford Handbook of Cognitive Psychology 206*.

Naumann, D., D. Frassinelli, and S. Schulte im Walde (2018). Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, New Orleans, LA, USA, pp. 76–85.

Pecher, D., I. Boot, and S. Van Dantzig (2011). Abstract Concepts. Sensory-Motor Grounding, Metaphors, and Beyond. *Psychology of Learning and Motivation – Advances in Research and Theory 54*, 217–248.

Pollock, L. (2018). Statistical and Methodological Problems with Concreteness and other Semantic Variables: A List Memory Experiment Case Study. *Behavior Research Methods 50*(3), 1198–1216.

Schäfer, R. and F. Bildhauer (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 486–493.

Turney, P., Y. Neuman, D. Assaf, and Y. Cohen (2011). Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, pp. 680–690.

Vigliocco, G., S.-T. Kousta, P. A. Della Rosa, D. P. Vinson, M. Tettamanti, J. T. Devlin, and S. F. Cappa (2014). The Neural Representation of Abstract Words: The Role of Emotion. *Cerebral Cortex 24*(7), 1767–1777.