Challenges for corpus annotation of copulative perception verbs

Alon Fishman
Digital Humanities and Social Sciences Hub
The Open University of Israel
Raanana, Israel 4353701
alonfishm@gmail.com

Relevant UniDive working groups: WG1

1 Introduction

Copulative perception verbs (CPVs) such as English *look (like)* and Hebrew *nir'e* 'look (like)' are characterized by taking a perceived object as their grammatical subject, and requiring a predicative or clausal complement (Rogers, 1973; Viberg, 1983, 2019; Gisborne, 2010; Poortvliet, 2018; Fishman, 2023; Melnik, In press). In the linguistics literature, they have received little attention relative to other perception verbs, especially in non-European languages. In the context of corpus annotation, they present interesting challenges and are currently covered inconsistently across languages.

The current study has the following aims: presenting an overview of CPVs, with a focus on commonalities and idiosyncrasies between languages; describing desiderata and challenges for annotation of CPVs in corpora; and reviewing the annotation of CPVs in currently available corpora. As a study of a syntactic and semantic phenomenon which is weakly covered in annotated corpora, this work is most relevant to Working Group 1. Since many of the forms occurring as CPVs have other senses as well, which could potentially be linked with their occurrences in corpora, this work could potentially be of interest to Working Group 2 as well. Finally, this work focuses on a relatively rare phenomenon, with subtly different expressions across languages, and thus may be of interest to Working Group 4.

2 Background

In the linguistic domain of perception, a fundamental distinction is drawn between experiencer-based (or subject-oriented) verbs and source-based (or object-oriented) verbs (Viberg, 1983). Experiencer-based verbs take a perceiver as their grammatical subject and refer to an experience of perception by that perceiver, whether volitional (e.g., *look (at)*, *listen*) or non-volitional (e.g., *see*, *hear*). Source-based verbs, on the other hand, take a perceived object as their grammatical subject and refer to a perceptual impression of that object. They may be further divided into CPVs (e.g., *look (like)*, *sound*), which require a predicate or clausal complement, and verbs which are predicates in themselves (e.g., *shimmer*, *buzz*) (Viberg, 2019).

Syntactically, CPVs typically take adjectival or adverbial complements, comparative complements, and clausal complements, with the latter also available in impersonal constructions; see (1)-(2). Semantically, many CPVs have two distinct meanings: an attributary meaning which attributes a property to a perceptual impression, and an evidential meaning which relates a proposition to a source of evidence; contrast (1a) with (1b) (Gisborne, 2010).

(1) He looks bad.

  a. 'His appearance is unattractive'. (attributary)

  b. 'Judging by his appearance, he is malicious'. (evidential)

(2) a. He looks like a linguist.

  b. He looks like/as if he's a linguist.

  c. It looks like/as if he's a linguist.

These two meanings are grammatically indistinguishable in most languages where CPVs have been studied in depth (e.g., Whitt, 2009; Gisborne, 2010; Poortvliet, 2018; Viberg, 2019; Staniewski and Gołębiowski, 2021). However, the two can be distinguished in Hebrew, where instances expressing attributary and evidential meanings take adverbial and adjectival complements, respectively; contrast (3a) with (3b) (Avineri, 2021; Fishman, 2023).

(3) Hebrew

  a. hem nir'im ra.
  they look.MPL badly
  'They look bad (= unattractive).'

  b. hem nir'im ra'im.
  they look.MPL bad.MPL
  'They look bad (= malicious).'

Similarly to the Hebrew alternation, Russian *vygljadit* 'look' takes both adjectival and adverbial complements, and Finnish CPVs take complements with both ablative and allative case; see (4)-(5). Following Fishman (2023), I use Distinctive Collexeme Analysis (Gries and Stefanowitsch, 2004) to explore whether these formal distinctions correspond to a semantic distinction, with the ultimate aim of establishing a taxonomy of CPV form-function pairings.

(4) Russian
On vygljadit plox-o/-im.
he looks.SG bad.ADV/SG.INS
'He looks bad.'

(5) Finnish
Tämä näyttää paha-ltä/-lle.
this looks.SG bad.ABL/ALL
'This looks bad.'

3 Corpus annotation

As a linguist seeking to explore the use of CPVs in corpora, my criteria for successful corpus annotation of CPVs are (i) reliable identification, (ii) generalization over different verbs within a language, and (iii) generalization over verbs across languages. These aims are far from trivial and may in fact prove to be impossible in certain cases. As noted above, CPVs may take a wide variety of complements, may occur with or without a logical subject, and may have multiple distinct meanings, all of which could potentially pose challenges for identifying CPVs.

Many of the forms which occur as CPVs also occur as other verb types (Viberg, 1983), which further complicates efforts to identify them. Moreover, such polysemies are not always systematic across verbs, even within a single language, let alone across languages. For example, English *look* occurs as an experiencer-based verb (e.g., *Look at this*), whereas *sound* occurs as a source-based predicate (e.g., *The alarm sounded*), and *smell* occurs as both (contrast *Smell this* with *The shower smells*). Perhaps the most extreme example of this challenge are verbs such as *feel*, which take many of the same types of complements when occurring as a CPV and as an experiencer-based verb; see (6)-(8).

(6) a. The bed feels cold.
b. I feel cold.

(7) a. The bed feels like a block of ice.
b. I feel like a block of ice.

(8) a. The bed feels like nobody slept here.
b. I feel like nobody slept here.

These challenges are exacerbated when trying to unify annotation cross-linguistically, due to the idiosyncracies of CPVs in different languages. For example, English CPVs take adjectival but not adverbial complements. In contrast, Polish CPVs (e.g., *pachnieć* 'smell') take adverbial but not adjectival complements (Staniewski and Gołębiowski, 2021). And as mentioned above, Hebrew CPVs take both adjectival and adverbial complements, with a different meaning for each.

UD guidelines address CPVs with adjectival complements under the heading of secondary predication, with the English example *He looked fantastic*. As the complement is obligatory, the guidelines advise that it should be attached as an `xcomp` of the verb, and this is upheld quite reliably in, e.g., the English Web Treebank, (Silveira et al., 2014), though less so in non-English treebanks. However, in treebanks for languages where CPVs take adverbial complements (e.g., Hebrew, Russian and Polish), the criterion of obligatoriness seems to be disregarded, and the complement is most often attached as an `advmod` of the verb. Similar inconsistencies arise in the ways clausal complements are attached to CPVs, between treebanks for different languages and even between treebanks for the same language.

I inspect the annotation of CPVs and their complements in treebanks available on GREW-MATCH, in the following languages: English, Hebrew, Russian and Finnish (and potentially Polish, Spanish and German, as time permits). I report inconsistencies in the annotation of each complement type, within and across treebanks. My hope is that this study, and any discussions it may give rise to, would both shed light on an understudied class of verbs within linguistics, and advance efforts to address a challenge in corpus annotation.

References

Bar Avineri. 2021. Alternating smell in Modern Hebrew. In Łukasz Jędrzejowski and Przemysław Staniewski, editors, *The linguistics of olfaction*. John Benjamins, Amsterdam, NL.

Alon Fishman. 2023. Hebrew copulative perception verbs. *Linguistics*, 61(4):997–1026.

Nikolas Gisborne. 2010. *The event structure of perception verbs*. Oxford University Press, Oxford, UK.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on "alternations". *International Journal of Corpus Linguistics*, 9:97—-130.

Nurit Melnik. In press. Copy raising reconsidered. *Journal of Language Modelling*.

Marjolein Poortvliet. 2018. *Perception and predication: A synchronic and diachronic analysis of Dutch descriptive perception verbs as evidential copular verbs*. Ph.D. thesis, University of Oxford.

Andrew D. Rogers. 1973. *Physical perception verbs in English: A study in lexical relatedness*. Ph.D. thesis, UCLA.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Przemysław Staniewski and Adam Gołębiowski. 2021. To what extent can source-based olfactory verbs be classified as copulas? In Łukasz Jędrzejowski and Przemysław Staniewski, editors, *The linguistics of olfaction*. John Benjamins, Amsterdam, NL.

Åke Viberg. 1983. The verbs of perception: A typological study. *Linguistics*, 21(1):123–162.

Åke Viberg. 2019. Phenomenon-based perception verbs in swedish from a typological and contrastive perspective. *Syntaxe et Sémantique*, 20:17–48.

Richard J. Whitt. 2009. Auditory evidentiality in english and german: The case of perception verbs. *Lingua*, 119:1083—-1095.