

Self, but not selfish: The interpretation of Hebrew self-anaphors

While it is well established that pronouns permit both bound-variable (sloppy) and coreferential (strict) interpretations, there is no consensus regarding the interpretive possibilities of self-anaphors. Traditionally, self-anaphors are assumed to allow only bound-variable readings [1]. This neat separation has been challenged by claims that self-anaphors, too, may receive a strict interpretation [2,3]. To our knowledge, McKillen (2016) [4] is the only work that has experimentally examined this question, suggesting English self-anaphors do permit strict readings in focus constructions. Its design, however, relied on explicit interpretive instructions and on contexts whose relevance to the interpretation of the target sentences was not made explicit to participants, leaving it unclear whether the acceptance found reflects true acceptability or task-induced bias. To clarify the interpretations available to self-anaphors and identify the sources of disagreement, the present study experimentally investigated how native speakers interpret Hebrew self-anaphors relative to pronouns in *only*-focus constructions.

Design and Materials. We crossed Type of Pronominal Element: Anaphor (A) vs. Pronoun (P) with Type of Reading: Sloppy (SL), Strict (ST) or Infelicitous (INF), yielding six conditions (1). 12 sets of structurally parallel, well-formed items were constructed (72 in total), each pairing a preamble that introduced a pronominal element under *only*-focus with a refutational response beginning with a denial (e.g., *aval ze lo naxon*, “but that’s not true”), thus forcing the intended reading. The preambles established clear focus-alternative sets distinguishing strict from sloppy readings. Pronouns, which permit both interpretations, served as a positive baseline, whereas an erroneous-focus response (Infelicitous) served as a negative baseline.

(1)

Preamble (A): *ha-me'araxat ha-ce'ira amra še-ba-mesiba etmol rak dan hicig et acmo.*

‘The young hostess said that at the party yesterday only Dan introduced himself.’

- a. ... *gam šlomit hiciga et acma.* ‘... Shlomit also introduced herself.’ (A_SL)
- b. ... *gam šlomit hiciga oto.* ‘... Shlomit also introduced him.’ (A_ST)
- c. ... *dan hicig gam et šlomit.* ‘... Dan also introduced Shlomit.’ (A_INF)

Preamble (P): *ha-me'araxat ha-ce'ira amra še-ba-mesiba etmol rak dan hicig et ha-ben šelo.*

‘The young hostess said that at the party yesterday only Dan introduced his son.’

- a. ... *gam šlomit hiciga et ha-ben šela.* ‘... Shlomit also introduced her son.’ (P_SL)
- b. ... *gam šlomit hiciga oto.* ‘... Shlomit also introduced him.’ (P_ST)
- c. ... *dan hicig gam et šlomit.* ‘... Dan also introduced Shlomit.’ (P_INF)

Methods. 123 native Hebrew speakers (aged 18-35; 21 excluded, leaving 102) rated, on a 7-point Likert scale, how natural each response was as a reaction to its preamble; the instructions emphasized judging the suitability of the response rather than the plausibility of the scenario. Filler items shared the preamble-response format at varying levels of acceptability. Items were distributed over six lists (Latin square design; 12 experimental items + 24 fillers each) and presented in a pseudo-randomized order, with two filler items intervening between any two experimental items, so that each participant saw two items per condition.

Results. Ordinal ratings were analyzed with a Cumulative Link Mixed Model (Pronominal Element × Reading; by-participant and by-item random intercepts). The interaction was significant ($\chi^2(2) = 34.66$, $p < .001$). Both pronouns and anaphors were near ceiling under the sloppy reading and near floor under the infelicitous reading, differing in neither; they diverged only under the strict reading, where anaphors were rated significantly lower than pronouns ($\beta = -1.48$, $p < .001$). Planned contrasts within anaphors confirmed that the strict reading was significantly less acceptable than the sloppy reading ($\beta = 3.13$, $p < .001$), yet remained significantly above the infelicitous baseline ($\beta = 1.30$, $p < .001$; Table 1, Figure 1). The intermediate strict-anaphor mean (Table 1) does not reflect uniform mild degradation: as the distribution of individual participant means shows (Figure 2), ratings span the entire scale, concentrated in the mid-range but with lower ratings more densely represented than higher ones, indicating substantial inter-speaker variability rather than a shared, moderate judgment.

	INF	SL	ST
Pronoun	2.28 (1.75)	6.43 (1.23)	5.46 (1.88)
Anaphor	2.57 (2.00)	6.60 (0.95)	3.88 (2.19)

Table 1. Mean acceptability (SD) by condition (INF = infelicitous, SL = sloppy, ST = strict).

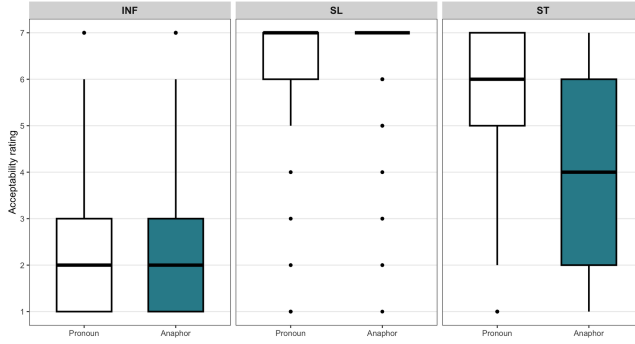


Figure 1

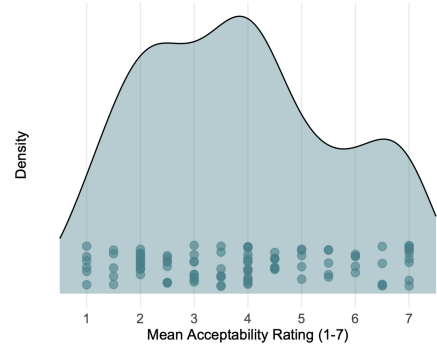


Figure 2

Figure 1. Acceptability ratings by Type of Pronominal Element (Pronoun, Anaphor) and Type of Reading (INF, SL, ST). **Figure 2.** Distribution of individual participant means in the strict-anaphor condition (n = 102).

Discussion. Two findings stand out. First, self-anaphors differ from pronouns in interpretation (not only in locality): the strict reading of the anaphor is significantly less acceptable than that of the pronoun, and than the anaphor's own sloppy reading. Second, this strict reading is nonetheless not entirely ruled out by the grammar, as it is rated significantly above the infelicitous baseline, thus available, if only to a limited and highly variable extent. We take both properties to follow from the internal structure of the anaphor. In the spirit of previous proposals [5,6], we suggest that self-anaphors have a complex structure, schematically [DP₁ [D +def] [NP [DP₂ pron] [N *ecem*- 'self']]], in which the pronominal core (DP₂) is embedded within the anaphor [7]. Under *only*, focus alternatives are formed by substituting the focused constituent with structurally no more complex expressions [8]. Copying the full anaphor (DP₁) into an alternative yields the sloppy reading, which the parser computes by default. To derive the strict (coreferential) reading, we suggest that the computational system applies Vehicle Change (VC) to the anaphor in the focus alternatives [9,10], reducing it to its pronominal core (DP₂) and thereby yielding the strict reading. Because VC is required for interpretation rather than convergence, it involves constructing a Reference Set that computes and compares the relevant focus alternatives and their corresponding sloppy and strict interpretations. Since constructing this Reference Set is optional rather than required for grammatical well-formedness, it is computationally costly, and its outcome is graded rather than categorical [11]. This account predicts exactly the profile we observe: a strict reading that is grammatically available yet significantly less acceptable than the sloppy one. It also predicts substantial inter-speaker variability, because speakers differ in their willingness or ability to perform this costly computation, resulting in participant mean acceptability ratings that span nearly the full scale (Figure 2) rather than clustering within a narrow range. Further support for the account comes from previous observations that when context or world knowledge independently disfavors the sloppy reading, the default interpretation no longer competes, making the strict reading readily accessible (e.g., *In the 1992 elections, Clinton voted for himself, and of course, Hillary did too*) [3].

[1] Reinhart & Reuland (1993) *Ling. Inq.* [2] Sag (1976) PhD diss., MIT. [3] Hestvik (1995) *Nat. Lang. Semant.* [4] McKillen (2016) PhD diss., McGill. [5] Faltz (1977) PhD diss., Berkeley. [6] Bassel (2024) PhD diss., HUJI. [7] Middleton (2020) PhD diss., UCL. [8] Katzir (2007) *Ling. & Phil.* [9] Fiengo & May (1994) MIT Press. [10] Abels (2022) Ms., UCL. [11] Reinhart (2006) MIT Press.