Franziska Kretzschmar* and Ingmar Brilmayer

Zooming in on agentivity: Experimental studies of DO-clefts in German

https://doi.org/10.1515/lingvan-2019-0069

Abstract: Despite the importance of the agent role for language grammar and processing, its definition and features are still controversially discussed in the literature on semantic roles. Moreover, diagnostic tests to dissociate agentive from non-agentive roles are typically applied with qualitative introspection data. We investigated whether quantitative acceptability ratings obtained with a well-established agentivity test, the DO-cleft, provide evidence for the feature-based prototype account of (Dowty, David R. 1991. Thematic proto-roles and argument selction. *Language* 67(3). 547–619) postulating that agentivity increases with the number of agentive features that a role subsumes. We used four different intransitive verb classes in German and collected acceptability judgements from non-expert native speakers of German. Our results show that sentence acceptability increases linearly with the number of agentive features and, hence, agentivity. Moreover, our findings confirm that sentience belongs to the group of proto-agent features. In summary, this suggests that a multidimensional account including a specific mechanism for role prototypicality (feature accumulation) successfully captures gradient acceptability clines. Quantitative acceptability estimates are a meaningful addition to linguistic theorizing.

Keywords: semantic roles, agent, experiencer, sentience, acceptability judgements, role decomposition, role prototypicality

1 Introduction

Agentivity is a core component of human cognition with ramifications in language grammar and processing (Bornkessel-Schlesewsky and Schlesewsky 2014; Spelke and Kinzler 2007). Accordingly, there exists a continuing debate on the adequate definition of semantic roles (Levin and Rappaport-Hovav 2005; Rissman and Majid 2019), with at least two open questions regarding the agent role. First, role definitions work with different assumptions about role composition and role properties or features, leading to numerous proposals ranging from atomic roles to generalised roles sometimes defined as feature clusters (see overview in Levin and Rappaport-Hovav 2005). Consequently, diagnostic tests for roles and their features vary depending on the kind of role definition as well as the kind and number of features under investigation, and it is obvious, therefore, that they may not elicit converging results. Although there is growing evidence that feature-based definitions of the agent role are superior to atomic ones (cf. Ackerman and Moore 2001) what exactly counts as an agent(ive) role or feature remains somewhat vague.

Second, agentivity tests are often based on subjective introspection, i.e. expert judgements, that lack a quantitative foundation in native speakers' intuitions about the acceptability of a sentence. This may contribute to the heterogeneous pattern in defining the agent role and suggests that test results may suffer from low generalisability to the respective speaker community. Intriguingly, while criticism of this kind has been formulated with respect to work in theoretical syntax (e.g., Dabrowska 2010; Gibson and Fedorenko 2010; Gibson et al. 2013; Linzen and Oseki, 2018; Sprouse et al. 2013), judgement-based quantitative investigations

***Corresponding author: Franziska Kretzschmar**, CRC 1252 "Prominence in Language", University of Cologne, Cologne, Germany, E-mail: franziska.kretzschmar@uni-koeln.de

Ingmar Brilmayer, CRC 1252 "Prominence in Language", University of Cologne, Cologne, Germany, E-mail: ingmar.brilmayer@uni-koeln.de

on role properties of the agent or other roles are still rare (cf. Ambridge et al. 2016; Kako 2006; Reisinger et al. 2015; Schlesinger 1992), and make little contact with the diagnostic tests used for role definition.

The goal of the present study is twofold. First, we compare the agent notions of two leading theoretical accounts of agentivity: Dowty (1991), proposing feature-based generalised proto-roles, and Jackendoff (1993, 2007), proposing generalised macroroles. We focus on Dowty's approach, specifically on the status of sentience as an agentive feature vis-à-vis the features volition and movement, as it is treated differently by the two accounts. Our findings extend existing empirical research on the psychological reality of Dowty's role feature system (e.g., Kako 2006). Second, we chose the abovementioned accounts because they differently use a well-established role diagnostic: the DO-cleft construction, consisting of a clefted verb phrase and a *wh*-clause containing the verb *do* (*what x did was* <verb phrase>). We conducted two acceptability judgement experiments to investigate the reliability of the DO-cleft as it has not yet been rigorously scrutinised in quantitative research designs.

2 DO-clefts and notions of agentivity

DO-clefts are a well-established test for role-semantic properties of the agent or actor role (Cruse 1973; Halliday 1968; Jackendoff 1993, 2007; Klima 1961; Lakoff 1966), which is the relevant role for the sets of intransitive verbs (i.e. verbs with a single argument) in our study. While the construction appears not to be limited to volitional agents, the range of agentive features it targets has not been accurately identified in the literature. For instance, in Cruse's (1973) feature-based approach to agentivity any one feature from his list of agentive features suffices to license the construction.

On another account, Jackendoff (2007) postulates two superordinate generalised roles on the macrorole tier in his framework, Actor and Undergoer, and utilises DO-clefts to test for the Actor role. Using examples such as (1) and (2), he provides introspective acceptability judgements to distinguish verbs that select the Actor role from those that don't (Jackendoff 2007: 197, 204):

(1) a. *What Bill did was throw the ball.*
b. *What the ball did was roll to the wall.*

(2) a. **What Bill did was own a VW.*
b. **What I did was see the tree.*

He assumes that the verbs in (1) select an Actor and are therefore acceptable in the DO-cleft. The verbs in (2) lack an Actor role, hence yielding unacceptable strings. Jackendoff does not define the Actor in terms of specific role properties, and, hence, is not explicit about the type of Actor that is acceptable in the DO-cleft. A comparison of the acceptable volitional, animate Actor argument in (1a) and the non-volitional, inanimate Actor argument in (1b) suggests that the DO-Cleft is insensitive to volition as a role feature (and role-related animacy information). By contrast, comparing (1b) with the unacceptable non-volitional experiencer subject in (2b) seems to support the view that the DO-cleft construction distinguishes moving agents from sentient ones.

This treatment of sentience corresponds to its exclusion from the inventory of agentive features in some feature-based agentivity accounts (e.g., Cruse 1973; Schlesinger 1992). One of the few to include sentience is Dowty's (1991) proto-role approach. He postulates two generalised proto-roles, Proto-Agent and Proto-Patient, being composed of bundles of entailments generated by the verb's meaning. Dowty (1991: 572) lists five features for the Proto-Agent role, which may occur in isolation or in combination: volition, sentience, causation, movement and independent existence. The first three features are crucial for our tested verb sets (cf. Section 3). In describing verbs like *nominate* or *murder*, Dowty (1991: 552, 607) states that a volitional agent acts volitionally and intentionally, and is necessarily sentient. A moving agent moves autonomously using its own source of energy, or otherwise, movement is a proto-patient property (Dowty 1991: 574). Dowty (1991: 554, 607)

attributes movement to any form of activity of the participant in question, including the subtle mental activity entailed by volitional perception verbs (e.g., *look at*), or verbs denoting bodily processes (e.g., *bleed*). A sentient agent is sentient of, i.e. able to perceive, mentally represent or evaluate, "the state or event denoted by the verb" (1991: 573), and occurs, inter alia, with subject experiencer verbs, i.e. emotion (e.g., *fear*), cognition (e.g., *know*) and perception (e.g., *see*) predicates.

There are three dimensions, relevant to our study, on which Dowty deviates from Jackendoff's concept of agentivity. First, Dowty assumes that there are clines of agentive roles depending on the number of agentive features these roles accumulate. These clines can be tied to agent prototypicality as follows. The agent prototype accumulates the highest number of agentive features, and an argument with a higher number of agentive features will be closer to the prototype (i.e. more agentive) and preferably selected as the subject of the verb (Dowty 1991: 576). In this way, role prototypicality positively correlates with the number of agentive features, while the kind of feature is secondary. Second, Dowty uses the DO-cleft indecisively: he cites one of Cruse's (1973) DO-cleft examples as a diagnostic for volition (Dowty 1991: 572) and uses the construction elsewhere as a dynamicity test to distinguish statives from dynamic predicates (Dowty 1979: 55; see also Smith 1999), but he never uses it to evaluate role prototypicality. Finally, Dowty's and Jackendoff's agentivity notions treat sentience verbs differently. According to Dowty (1991: 573), verbs such as *see* in (2b) entail the proto-agent feature sentience, while *roll* in (1b) entails movement for the subject argument. Following his role prototypicality hypothesis above, *see* and *roll* are equally remote from the prototype, as both entail only one agentive feature and because each feature counts equally towards agentivity. Hence, contra Jackendoff (2007), examples (1b) and (2b) are not expected to exhibit different acceptability behaviour in Dowty's framework.

In summary, there is good evidence that the DO-cleft is sensitive to agentive features, but introspection data fail to answer which particular set of agentive features restricts its use. Additionally, while there is empirical support for Dowty's feature-based proto-role approach (Kako 2006; Reisinger et al. 2015), his operationalisation of role prototypicality, i.e. feature accumulation, has been investigated less often. Recently, Kretzschmar et al. (2019) further investigated feature accumulation in the DO-cleft construction in German using quantitative acceptability ratings. They tested five classes of transitive verbs with different numbers of agentive features following Dowty (1991), as given in Table 1.

They found that volitional perception verbs (WATCH) were rated best, followed by sentience verbs (SEE, HATE, KNOW), which were rated better than ascription verbs (EXHIBIT; cf. Table 1). Because acceptability increased linearly with feature number, the authors argued that the empirical cline is fully compatible with Dowty's assumption that agentivity increases with the number of agentive features. By contrast, Jackendoff's (2007) account is challenged by the intermediate position of the sentience verbs, because it predicts that they should be treated like ascription verbs, both lacking the Actor macrorole. Finally, the acceptability cline challenges Dowty's (1979) use of DO-clefts as a dynamicity test, because the stative sentience verbs (SEE, HATE, KNOW) were rated better than the stative ascription verbs (EXHIBIT).

While Kretzschmar et al.'s (2019) acceptability cline challenges both Dowty's and Jackendoff's use of DO-clefts, it is compatible with Dowty's concept of role prototypicality, specifically the feature accumulation method, and his treatment of sentience as an agentive feature. However, the experiment cannot uncover whether the kind of agentive feature is also relevant. Recall that movement and sentience have yielded

Table 1: Verb classes, feature analysis and empirical acceptability cline from Kretzschmar et al. (2019).

Verb classes (example verbs)	**WATCH (*look at, watch*)**		**SEE (*see, hear*)**	**HATE (*love, hate*)**	**KNOW (*know, believe*)**		**EXHIBIT (*exhibit, have*)**
Feature analysis following Dowty (1991)	volition sentience movement		sentience	sentience	sentience		–
Mean acceptability rating	4.16	>	3.72	3.52	3.52	>	3.08

Note: Mean ratings are derived from a 6-point scale (1-very unacceptable to 6-very acceptable). '>' between mean ratings indicates significantly higher ratings for the verb class occurring to the left vs. the one occurring to the right.

opposite expert judgements in theoretical accounts, as argued for in (1b) vs. (2b) above. In the present experiments, we extend Kretzschmar et al.'s experimental design to include verb classes with sentience or movement as the only agentive feature.

3 Experiment 1: DO-clefts in supportive context

Following Kretzschmar et al.'s (2019) design, we included verb classes with varying numbers of agentive features, ranging from three to zero features. Additionally, we manipulated the kind of feature entailed by verb classes with an equal number of features. Table 2 illustrates the four intransitive verb classes.

Volitional action verbs (WORK) select three agentive features. FEAR and SWEAT select only one feature, either sentience or movement. We used emotion predicates in the FEAR class, because they are the only sentience verbs that can be used intransitively in German. The SWEAT class includes process verbs denoting non-volitional body processes in German (see Rosen 1984: 64; Dowty 1991: 607 or Perlmutter 1978: 162 for discussion). Verbs in the GLITTER class select none of the tested agentive features and denote stimulus appearances and emissions (e.g., Perlmutter 1978: 163). Despite the stable subject selection for these verbs in German, the number of features still indicates role prototypicality and feature accumulation (see also Ackerman and Moore 2001 who extend Dowty's framework to intransitive verbs). This is important because feature accumulation has been shown to play a role for DO-clefts with transitive verbs, as shown in Section 2. Hence, we predict that if role prototypicality holds irrespective of the (in)transitivity of the verb, we should find effects of the number of agentive features also in the present experiments, i.e. we should replicate the pattern reported in Kretzschmar et al. (2019). Specifically, the WORK class is predicted to be rated better than either FEAR or SWEAT, while GLITTER is predicted to be rated worst. The comparison between FEAR and SWEAT verbs additionally allows us to investigate the influence of the kind of feature when the number of features is held constant. This is stated as prediction one in Table 3.

From Jackendoff (2007), we extrapolate the prediction that WORK and SWEAT should cluster against FEAR and GLITTER in receiving better ratings (prediction 2).

Table 2: List of verb classes with individual verbs and agentive features.

	WORK	**FEAR**	**SWEAT**	**GLITTER**
Selected verbs	*arbeiten* 'work' *tanzen* 'dance' *tratschen* 'gossip' *flüstern* 'whisper' *turnen* 'do gymnastics' *reden* 'talk'	*bangen* 'fear' *trauern* 'mourn' *zweifeln* 'doubt' *leiden* 'suffer' *frieren* 'be cold' *staunen* 'be astonished'	*schwitzen* 'sweat' *niesen* 'sneeze' *zittern* 'shiver' *husten* 'cough' *bluten* 'bleed' *stottern* 'stutter'	*glitzern* 'glitter' *schimmern* 'glisten' *stinken* 'stink' *müffeln* 'smell musty' *leuchten* 'glow' *glänzen* 'glisten'
Agentive features	volition sentience movement	sentience	movement	–

Table 3: Overview of predictions pertaining to possible acceptability clines.

Prediction source	**Predicted acceptability cline**
1. Feature accumulation (Dowty 1991)	WORK > FEAR, SWEAT > GLITTER
2. Actor role (Jackendoff 2007)	WORK, SWEAT > FEAR, GLITTER

Note: a > b : a should be significantly rated better than b; a, b: a and b should receive indistinguishable ratings.

3.1 Method

3.1.1 Participants

Fifty-nine students (39 female, 2 unclassified, mean age: 22.89 years, SD: 7.07) participated in this study voluntarily. All were monolingual native speakers of German. One further participant was excluded from analysis because of not following the task instruction.

3.1.2 Stimuli

Items were constructed following a one-factorial design with four levels for the factor verb class (WORK vs. FEAR vs. SWEAT vs. GLITTER). There were six different atelic verbs per verb class (see Table 2) and three lexically different items per verb, resulting in a total of 18 items per verb class. To increase the naturalness of the experimental items, we constructed 2-sentence paragraphs (see Table 4).

An item included a lead-in sentence always introducing a non-specific group of humans quantified by *viele* 'many'. In the target sentence, the group of humans was referred to with a personal pronoun (*sie* 'they') and the DO-cleft construction was followed by a subordinate clause that ended the paragraph in a plausible way. We constructed six negative control items of similar structure, but with inanimate referents as subject that served as patient argument of the clefted passive clause. Since volition is ruled out for passive subjects in German, these negative control items should be completely unacceptable. Negative control items and DO-clefts with WORK verbs should in principle allow participants to use the end points of the rating scale.

Table 4: Example stimuli for Experiment 1.

Verb class	Example item
WORK	*Trotz der schlechten Ausbeute waren viele Bergbauarbeiter unter Tage. Was sie taten, war zu arbeiten, obwohl die Prognosen über den Kohlebestand negativ aussahen.* Despite the poor yield, many mine workers were underground. What they did was work, although the future prospects about the coal stock seemed negative.
FEAR	*Wegen des schweren Unfalls waren viele Angehörige besorgt. Was sie taten, war zu bangen, auch wenn die meisten Opfer außer Lebensgefahr waren.* Because of the severe accident many relatives were worried. What they did was fear, even though most of the victims were no longer in peril.
SWEAT	*Wegen der defekten Klimaanlage im Zug waren viele Passagiere gereizt. Was sie taten, war zu schwitzen, weil selbst Deo nicht mehr wirkte.* Because of the broken AC on the train, many passengers were short-tempered. What they did was sweat, because even deodorant didn't work anymore.
GLITTER	*Wegen der starken Sonneneinstrahlung waren viele Säuglinge gut eingecremt. Was sie taten, war zu glänzen, weil die Mütter zu viel Creme verwendet hatten.* Because of the strong insolation, many infants were creamed on well. What they did was glisten, because the mothers had used too much lotion.
Negative control (NC)	*In der Wäscherei waren viele Hemden zerknittert. Was sie taten, war von den Angestellten eilig gebügelt zu werden, damit die Kunden ihre Wäsche bald abholen konnten.* At the laundry, many shirts were crinkled. What they did was being hastily ironed by the employees, so the clients could soon pick up their clothing.

Note: The DO-cleft construction is underlined in the German originals and English translations, but was not underlined for participants.

3.1.3 Procedure and analysis

Using a Latin Square design, items were distributed across three experimental lists in a fully balanced way, so that each participant saw six lexically different items in each verb class (one item per verb), and no item more

than once within a list. Each list contained 24 critical sentences and six ungrammatical negative control items (identical across lists), which is well below the recommendation of 100 items per list in order to minimise effects of fatigue or strategic responses (Sprouse et al. 2013). Each list was presented in two pseudorandomised orders.

Using a paper-and-pencil questionnaire, we randomly assigned each participant to one of the experimental lists and asked them to rate each paragraph as a whole for acceptability, using a 6-point Likert scale ranging from very acceptable to very unacceptable. Rating categories were labelled with letters of neutral value (A–F) and, to facilitate easy responses, emoticons additionally helped participants to determine the degree of (un)acceptability.

Prior to analysis, we excluded missing or ambiguous responses (0.1 % of all responses) and recoded the response categories so that A (very acceptable) corresponded to 6 and F (very unacceptable) to 1. We used multi-level cumulative logit regression (Agresti 2002; Bürkner and Vuorre 2018) to account for the ordinal scale of our response variable and to avoid inflated Type I and Type II errors and distorted estimates of effect size (Liddell and Kruschke 2018). The analysis was performed in R (version 3.4.1, R Development Core Team 2017) with the package ordinal (Christensen 2015). We fitted a model with verb class as fixed effect and participants and items as crossed random intercepts. As noted by a reviewer, ratings may be confounded by two additional factors: a verb's propensity to also occur in transitive frames and the agentivity potential of the noun material selected as antecedent of the subject in the critical DO-cleft sentences. Therefore, we first calculated the number of argument structures per verb based on a corpus study[1] and then modelled it as a co-variate (Sassenhagen and Alday 2016).

Next, we conducted an additional rating study with 60 participants (49 females; mean age: 23 years, SD: 3.72) who voluntarily rated the noun material for its agentivity potential. None of them participated in this or the second experiment. We used the task instruction from a previous study that also investigated the agentivity potential of German nouns (Frenzel et al. 2015), but kept the scale identical to the current main experiment. Participants rated each noun according to whether it is a good or bad event instigator on the 6-point scale described above (A-good event instigator, F-bad event instigator). As argued by Frenzel and colleagues, this task instruction targets "participants' intuitions regarding relatively prototypical" agents (Frenzel et al. 2015: 6). After converting the response categories to a numerical format (A = 6, F = 1), we calculated mean values per verb and item for each noun and included it as a second co-variate in the statistical model.

We then tested for the inclusion of the by-participants random slope for verb class, which, however, led to a convergence error, so we report the intercept model below (cf. Table 5). As we assumed that verb classes should be ordinally ranked for agentivity, the verb class factor was used with forward-difference coding, which compares the mean rating for one level of the verb class factor with the mean rating of the immediate next level. We implemented it so as to test pairwise contrasts from the most agentive to the least agentive verb class: WORK vs. FEAR, FEAR vs. SWEAT, SWEAT vs. GLITTER. Negative control items were not statistically analysed, because our predictions only concerned the relative difference between the four critical verb classes.

3.1.4 Results

Figure 1 illustrates the mean values and distribution of acceptability ratings, including negative control items for completeness. Table 5 presents the statistical results.

Overall, all critical verb classes are rated as very acceptable. Verbs in the WORK class are significantly better rated than in the FEAR class. Ratings for FEAR and SWEAT verbs were statistically indistinguishable. GLITTER verbs were rated significantly worse than all other classes.

The number of argument structures per verb and the agentivity potential of the noun antecedent for the subject pronoun did not reveal a reliable influence. This pattern is exclusively compatible with prediction 1 in Table 3.

1 See supplementary materials: https://osf.io/bw3sh/?view_only=e5bbc0705b36430986e6e65ba4cc2276.

Table 5: Results of the best fitting and converged regression model for Experiment 1.

Model summary			
	Number of parameters	**AIC**	**logLik**
Intercept model	12	3862.4	−1919.2

Fixed effect	**Estimate**	**Standard error**	***z*-value**	***p*-value**
WORK vs. FEAR	0.625	0.297	2.103	<0.04
FEAR vs. SWEAT	0.153	0.209	0.733	0.463
SWEAT vs. GLITTER	1.386	0.209	6.627	<0.001
Number of argument structures	0.104	0.070	1.489	0.136
Subject agentivity	0.135	0.072	1.885	0.059

Threshold coefficients			
Threshold	**Estimate**	**Standard error**	***z*-value**
1|2	−3.978	0.350	−11.359
2|3	−2.104	0.331	−6.351
3|4	0.129	0.326	0.396
4|5	2.047	0.331	6.168
5|6	3.759	0.346	10.860

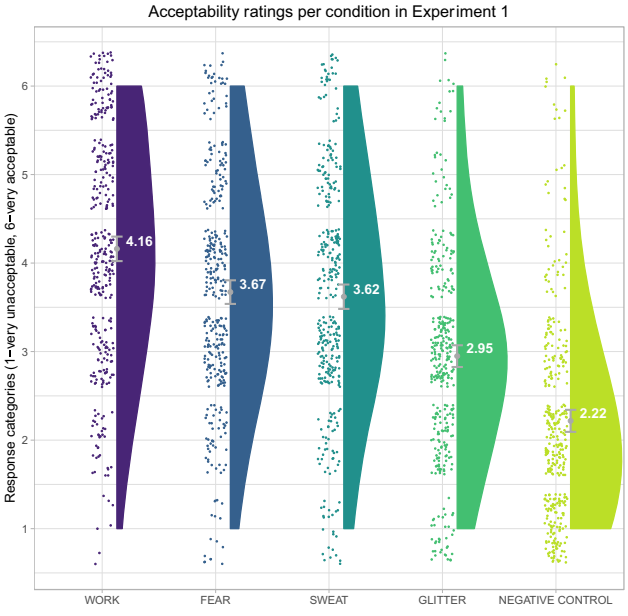

Figure 1: Experiment 1: Mean values and distribution of acceptability ratings. Error bars represent 95% confidence intervals.

3.2 Discussion

Experiment 1 revealed that acceptability increased with the increasing number of agentive entailments of the critical verbs, while not distinguishing between sentience or movement when occurring in isolation (WORK > FEAR, SWEAT > GLITTER). The gradient acceptability data indicate that the DO-cleft construction is sensitive to agentive features and their accumulation. Verbs entailing the highest number of agentive features, which can be considered as selecting prototypical agents, show the strongest advantage, but verbs entailing only sentience or movement also fare better than verbs selecting none of the tested agentive role features.

An important difference between Experiment 1 and the usage of the DO-cleft in the theoretical literature is that here the DO-cleft was embedded in a supportive, associative context, whereas the theoretical accounts presented in Section 2 usually present DO-clefts in isolation to test for verb differences. We address this possible confound in Experiment 2.

4 Experiment 2: DO-clefts without supportive context

In order to rule out that the findings from Experiment 1 were due to participants rating the fit between context sentence and DO-cleft, we presented DO-cleft sentences in isolation.

4.1 Method

4.1.1 Participants

Fifty students from the University of Cologne (31 females, mean age: 23.2 years, SD: 3.19) participated in this study voluntarily. All were monolingual native speakers of German. None of them participated in Experiment 1.

4.1.2 Stimuli

We used the same stimuli as in Experiment 1 (see Table 6) but removed the lead-in sentence and the subordinate sentence following the DO-cleft construction. In order to keep referential relations between the subject in the DO-cleft and the clefted verb comparable to Experiment 1, we replaced the personal pronoun with its antecedent noun phrase (placed in the lead-in sentence in Experiment 1).

4.1.3 Procedure and analysis

Procedure and analysis protocols were identical to Experiment 1. Participants' task instructions now mentioned the single sentences, instead of paragraphs. There were no invalid responses in Experiment 2. The model including a random slope for verb class by-participants converged and showed a better fit than the intercept-only model.

Table 6: Example stimuli for Experiment 2.

Verb class	Example item
WORK	*Was viele Bergbauarbeiter taten, war zu arbeiten.* What many mine workers did was work.
FEAR	*Was viele Angehörige taten, war zu bangen.* What many relatives did was fear.
SWEAT	*Was viele Passagiere taten, war zu schwitzen.* What many passengers did was sweat.
GLITTER	*Was viele Säuglinge taten, war zu glänzen.* What many infants did was glisten.
Negative control (NC)	*Was viele Hemden taten, war von den Angestellten eilig gebügelt zu werden.* What many shirts did was being hastily ironed by the employees.

4.1.4 Results

Figure 2 illustrates the mean values and distribution of acceptability ratings, including negative control items for completeness. Table 7 provides the statistical results.

The pattern of results resembles Experiment 1. The WORK class is significantly better rated than the FEAR and SWEAT classes, which did not differ from each other. GLITTER verbs were rated significantly worse than all other classes. The number of argument structures per verb and subject agentivity had no reliable influence.

4.2 Discussion

Experiment 2 served to validate the results from Experiment 1 regarding the question of whether associative context may influence acceptability ratings for agentivity features in DO-clefts. As the current experiment replicated the cline in Experiment 1 (WORK > FEAR, SWEAT > GLITTER), we can conclude that it is the verbs' fit into the DO-cleft construction that is being evaluated, not the overall fit into the preceding context. Nonetheless, ratings were somewhat lower in absolute terms compared with Experiment 1. This may indicate that the associative context in Experiment 1 primed the lexical content of the DO-cleft sentence, thereby increasing its acceptability across conditions.

5 General discussion

The present study experimentally tested a feature-based analysis of the agent role and the empirical reliability of the well-established DO-cleft test for agentivity. We found a cline where acceptability increases with the number of agentive features as defined in Dowty (1991), but not as a function of whether a verb selects the Actor role as defined in Jackendoff (1993, 2007). The cline occurred both with and without supportive context and is summarised in Table 8.

These findings support our prediction derived from Dowty's feature accumulation account (cf. Table 3) and mirror the experimental findings by Kretzschmar et al. (2019), using the same task and rating scale and

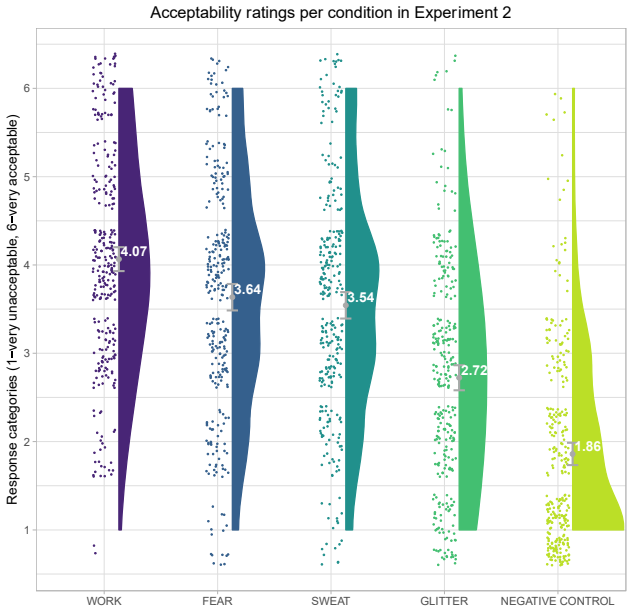

Figure 2: Experiment 2: Mean values and distribution of acceptability ratings. Error bars represent 95% confidence intervals.

Table 7: Results of the model comparison and summary of the best fitting regression model for Experiment 2.

Model comparison					
	Number of parameters	**AIC**	**logLik**	**LR.stat**	**df**	***p*-value**
Intercept model	12	3088.9	−1532.5			
Random slope model	21	3026.3	−1492.2	80.55	9	<0.001

Model summary				
Fixed effect	**Estimate**	**Standard error**	***z*-value**	***p*-value**
WORK vs. FEAR	1.102	0.359	3.071	<0.003
FEAR vs. SWEAT	0.228	0.249	0.918	0.358
SWEAT vs. GLITTER	2.155	0.323	6.653	<0.001
Number of argument structures	−0.007	0.081	−0.087	0.930
Subject agentivity	−0.078	0.141	−0.554	0.579

Threshold coefficients			
Threshold	**Estimate**	**Standard error**	***z*-value**
1|2	−4.921	0.809	−6.081
2|3	−2.952	0.798	−3.699
3|4	−0.709	0.794	−0.893
4|5	2.152	0.798	2.695
5|6	4.163	0.811	5.130

Table 8: Acceptability cline in the present experiments and its relation to the agentivity notions of Dowty (1991) and Jackendoff (2007).

Acceptability cline	**WORK**	**>**	**FEAR**	**SWEAT**	**>**	**GLITTER**
Dowty's agentive features	volition sentience movement		sentience	movement		–
Jackendoff's Actor notion	Actor		Non-Actor	Actor		Non-Actor

Note: '>' indicates significantly better ratings for the verb class occurring to the left vs. the one occurring to the right.

showing that graded acceptability ratings for agentive arguments in German positively correlate with the number of agentive features (cf. Table 1).

Yet, although the use of the same rating scale facilitates comparison across the experiments reported here and in Kretzschmar et al. (2019), we cannot conclude at present that our findings are truly independent of the scale format. For instance, scales with an even number of response categories avoid the interpretive ambiguity of the midpoint associated with an uneven number of categories, but tend to highlight responses near the endpoints of a scale (Weijters et al. 2010). Future research should address whether different scale formats induce comparable acceptability differences for role typicality, as has been reported for syntactic phenomena (e.g., Häussler and Juzek 2017; Weskott and Fanselow 2011).

More generally, our study focussed on two research questions: (i) Can we find further empirical support for Dowty's multidimensional notion of agentivity, i.e. a feature-based role definition with a specific feature accumulation method? And specifically, what is the status of sentience as an agentive feature vis-à-vis volition and movement? (ii) Are acceptability judgements about the DO-cleft that are based on subjective expert introspection supported by quantitative acceptability ratings?

With regard to question (i), our data, together with the findings of Kretzschmar et al. (2019) for transitive verbs, are compatible with Dowty's (1991) list of equally ranked Proto-Agent features, which includes sentience

in addition to, inter alia, movement and volition. Even though the experiencer of a sentience verb is not as close to the agent prototype as the role of the subject argument of volitional perception (e.g., *watch*) or action verbs (e.g., *work*), it still qualifies as an agentive role. This supports further theoretical accounts postulating that the experiencer is an agentive role (Rapp 1997; Schlesinger 1992). Also, the acceptability cline in Table 8 supports Dowty's feature accumulation as a mechanism underlying effects of role prototypicality. Types of agentive roles that are closer to the prototype, i.e. accumulate more features, are more acceptable in the DO-cleft.

By contrast, feature-based approaches without a specified mechanism for feature interaction (e.g., accumulation), such as Cruse's (1973), cannot explain our quantitative findings. In order to capture the entire empirical cline, any list of agentive features must be accompanied by such a mechanism. Moreover, accounts that do not distinguish between different types of generalised roles such as Jackendoff's (2007) Actor macrorole are too coarse to explain the graded pattern. Recall that for Jackendoff (2007) sentience verbs are unacceptable in the DO-cleft construction, whereas verbs with volitional or moving actor arguments are equally acceptable (see Section 2 and the second prediction in Table 3). Both assumptions are disconfirmed by the current experiments: verbs entailing volition, movement and sentience are more acceptable than verbs entailing only movement; the latter also do not differ from sentience verbs.

The finding that graded acceptability depends on the number of agentive role features also bears on the issue of appropriate role definitions more generally. Even though semantic roles may not be appropriately defined "in terms of necessary or sufficient conditions" (Rissman and Majid 2019: 1; cf. Levin and Rappaport-Hovav 2005), there is growing evidence for the empirical relevance of role features as defined by Dowty (e.g., Kako 2006; Reisinger et al. 2015; White et al. 2017). However, the existing empirical research has not specifically focussed on Dowty's feature accumulation method that also distinguishes his framework from other feature-based notions of agentivity (e.g., Cruse 1973). This is important because experimental evidence for the prototype structure of semantic roles, including typicality effects, has recently been argued to be rare (Rissman and Majid 2019: 13–14). By investigating Dowty's feature accumulation account, the present study lends new empirical support to feature accumulation as a mechanism underlying (proto-)typicality effects.

As regards question (ii), our results demonstrate that the DO-cleft is a reliable and sensitive test for agentivity, but also reveal partial divergence from qualitative introspection data in the theoretical literature. Importantly, our quantitative data show a more fine-grained acceptability cline than previously assumed. The test appears sensitive to the number of role features rather than the kind of feature, yielding graded acceptability. Hence, our findings suggest that the test is applicable to agentive role features rather than to predicate properties such as dynamicity. Overall, this supports previous claims that expert judgements do not yield entirely inaccurate acceptability contrasts (e.g., Sprouse et al. 2013). However, in line with others (Gibson et al. 2013), we find that quantitative estimates of effect size are a meaningful addition to linguistic theorizing – pointing to possible theoretical consequences for the definition of the agent role.

To conclude, we provide quantitative evidence that, contrary to introspection data, empirical agentivity profiles indicated by the DO-cleft test are not binary in distinguishing agents proper from all kinds of non-agent roles. Rather, they are graded. This gradience can be successfully captured with a multidimensional account of the agent role that includes an inventory of interacting features and sentience as one of them.

Acknowledgments: The research reported here was funded by the German Research Foundation (DFG) as part of the CRC 1252 "Prominence in Language" (Project ID 281511265 – SFB 1252). We thank Beatrice Primus for valuable comments on a previous version of the manuscript and our student assistants for their help during data collection.

References

Ackerman, Farrell & John Moore. 2001. *Proto-properties and grammatical encoding: A correspondence theory of argument selection*. Stanford: CSLI Publications.

Agresti, Alan. 2002. *Categorical data analysis*. Hoboken: John Wiley and Sons.

Ambridge, Ben, Amy Bidgood, Julian M. Pine, Caroline F. Rowland & Daniel Freudenthal. 2016. Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science* 40. 1435–1459.

Bornkessel-Schlesewsky, Ina & Matthias Schlesewsky. 2014. Competition in argument interpretation: Evidence from the neurobiology of language. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 107–126. Oxford: Oxford University Press.

Bürkner, Paul-Christian & Matti Vuorre. 2018. Ordinal regression models in psychological research: A tutorial. *psyarxiv*. https://doi.org/10.31234/osf.io/x8swp.

Christensen, Rune & Haubo Bojesen. 2015. Ordinal – Regression models for ordinal data. R package version 2015.6.28. Source: http://www.cran.r-project.org/package=ordinal/.

Cruse, D. Alan. 1973. Some thoughts on agentivity. *Journal of Linguistics* 9(1). 11–23.

Dabrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27. 1–23.

Dowty, David R. 1979. *Word meaning and Montague Grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ. (Studies in Linguistics and Philosophy)*. Dordrecht, Holland: Reidel.

Dowty, David R. 1991. Thematic proto-roles and argument selection. *Language* 67(3). 547–619.

Frenzel, Sabine, Matthias Schlesewsky & Ina Bornkessel-Schlesewsky. 2015. Two routes to actorhood: Lexicalized potency to act and identification of the actor role. *Frontiers in Psychology* 6. 1.

Gibson, Edward & Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14(6). 233–234.

Gibson, Edward, Steven T. Piantadosi & Evelina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida. *Language and Cognitive Processes* 28(3). 229–240.

Halliday, Michael Alexander Kirkwood. 1968. Notes on transitivity and theme in English: Part 3. *Journal of Linguistics* 4(2). 179–215.

Häussler, Jana & Tom Juzek. 2017. Hot topics surrounding acceptability judgement tasks. In San Featherston, Robin Hörnig, Reinhild Steinberg, Birgit Umbreit & Jennifer Wallis (eds.), *Proceedings of linguistic evidence 2016. Empirical, theoretical, and computational perspectives*, 1–21. Tübingen: University of Tübingen. https://doi.org/10.15496/publikation-19039.

Jackendoff, Ray. 1993. The combinatorial structure of thought: The family of causative concepts. In Eric Reuland & Werner Abraham (eds.), *Knowledge and language, Vol. II, Lexical and conceptual structure*, 31–49. Dordrecht: Kluwer Academic Publishers.

Jackendoff, Ray. 2007. *Language, consciousness, culture: Essays on mental structure*. Cambridge, Mass: MIT Press.

Kako, Edward. 2006. Thematic role properties of subjects and objects. *Cognition* 101. 1–42.

Klima, Edward S. 1961. Structure at the lexical level and its implications for transfer grammar. In *International conference on machine translation of languages and applied language analysis*, 98–108.

Kretzschmar, Franziska, Tim Graf, Markus Philipp & Beatrice Primus. 2019. An empirical investigation of agent prototypicality and agent prominence in German. In Anja Gattnar, Robin Hörnig & Melanie Störzer (eds.), *Online Proceedings of linguistic evidence 2018 – experimental data drives linguistic theory*, 101–123. Tübingen: University of Tübingen Press.

Lakoff, George. 1966. Stative adjectives and verbs in English. In Anthony G. Oettinger (ed.), *Mathematical linguistics and automatic translation report no. NSF-17 to the National Science Foundation*, I–1–I–16. Cambridge, Mass: Harvard University Computation Laboratory.

Levin, Beth & Malka Rappaport-Hovav. 2005. *Argument realization*. Cambridge: Cambridge University Press.

Liddell, Torrin M. & John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79. 328–348.

Linzen, Tal & Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics* 3(1). 1–25.

Perlmutter, David M. 1978. Impersonal passives and the unaccusative hypothesis. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 38. 157–189.

R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: The R Foundation for Statistical Computing. https://www.R-project.org/.

Rapp, Irene. 1997. *Partizipien und semantische Struktur. Zu passivischen Konstruktionen mit dem 3. Status (Studien zur deutschen Grammatik 54)*. Tübingen: Stauffenburg Verlag.

Reisinger, Drew, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins & Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics* 3. 475–488.

Rissman, Lilia & Asifa Majid. 2019. Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*. 1–20.

Rosen, Carol. 1984. The interface between semantic roles and initial grammatical relations. In David M. Perlmutter & Carol Rosen (eds.), *Studies in relational grammar 2*, 38–80. Chicago: University of Chicago Press.

Sassenhagen, Jona & Phillip M. Alday. 2016. A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language* 162. 42–45.

Schlesinger, Izchak M. 1992. The experiencer as an agent. *Journal of Memory and Language* 31. 315–332.

Smith, Carlota S. 1999. States or events? *Linguistics and Philosophy* 22(5). 479–508.

Spelke, Elizabeth S. & Katherine D. Kinzler. 2007. Core knowledge. *Developmental Science* 10(1). 89–96.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248.

Weijters, Bert, Elke Cabooter & Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing* 27. 236–247.
Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87. 249–273.
White, Aaron S., Kyle Rawlins & Benjamin Van Durme. 2017. The semantic proto-role linking model. *EACL* 2017. 92–98.