Embedded questions:
Evidence in a decision-theoretic paradigm for 'surprise' & 'agree'

Lea Fricke[1], Emilie Destruel,[2] Edgar Onea[1] & Malte Zimmermann[3]
[1]Graz University, [2]University of Iowa, [3]Potsdam University

In this paper, we present new cross-linguistic empirical evidence on the interpretation of questions, specifically those embedded under the two verbs *surprise* and *agree,* in German and English. Our study contributes to the long-standing debate between introspection and experimental findings on what are the (exhaustive) readings that speakers actually derive in such embeddings. In our experiment, participants were confronted with a decision problem involving all potential exhaustive readings, for which we created probabilistic models of the participants' beliefs in a Bayesian analysis. Our results align with prior empirical evidence: embedded questions under *surprise* and *agree* are accepted under weakest readings.

1 Introduction

The present paper aims at answering the following question: What (exhaustive) readings associate with embedded questions and to what extent do speakers derive them? Here, we focus specifically on two embedding verbs, *surprise* and *agree*, illustrated in (1) and (2),

(1) Ali was **surprised** who danced at the party.
(2) Kim and Ali **agree** on who of the dancers lack talent.

Past theoretical literature discusses different readings that arguably exist for these two verbs, which are summarized in Table 1 and 2. Readings are ordered based on their logical strength, the top reading being the strongest. For *surprise*, the general assumption is that (1) is true under a weakly exhaustive reading (**WE**). This WE reading has two manifestations following Lahiri (2002), who claims this verb is only optionally distributive. Others have argued for strong exhaustivity (**SE**). Under *agree*, the question meaning characterizes the extent to which the attitude holders' (AH) beliefs must align. Incompletely aligned beliefs (**IA**) served as a negative baseline in our study.

Table 1
überraschen/surprise Q (A+ = A has Q-property; A- = A does not have Q-property)

Reading	Mental state of Attitude Holder	Facts in the World
WE$_{\text{distributive}}$ (Berman 1991)	Expectation: A-, B-, C-	A+, B+, C+, D-, E-
WE$_{\text{non-distributive}}$ (Lahiri 2002)	Expectation: A-, B-	
SE (Klinedinst & Rothschild 1999)	Expectation: A+, B+, C+, D+, E+	

Recent experimental evidence is partially at odds with some of these theoretical claims, suggesting that sentences like (1) and (2) are indeed true under the weakest interpretations, i.e., **SE** and **CA** (Cremers & Chemla 2017, Chemla & George 2016). However, the status of the readings detected in these experiments is unclear: Since the experiments used truth-value judgment tasks, the communicative reliability and robustness of the accepted readings remains unknown. Our study aims at extending this empirical line of research from a cross-linguistic perspective by testing the availability of pragmatically robust and reliable readings only.

Table 2
sich einig sein/agree Q (A+ = A has Q-property, A- = A lacks Q-property, A? = uncertain whether A has Q-property.)

Reading	Mental states of Attitude Holder 1 & Attitude Holder 2
CA+ (complete alignment) (Kratzer 2006)	Both AHs believe: A +, B+, C+, D-, E-
CASU (complete alignment, same uncertainties) (Beck & Rullmann 1999)	Both AHs believe: A +, B+, C+, D?, E?
CA (complete alignment of positive belief) (Lahiri 2002)	Both AHs believe: A+, B+, C+, D- AH1 believes: E-; AH2 believes: E?
IA (incomplete alignments)	Both AHs believe: A+, B+, C+, D- AH1 believes: E+; AH2 believes: E-

2 Experiment

2.1 Participants

A total of 24 native speakers of German (mostly Austrian German) were tested, namely 17 females and 7 males between the ages of 20 and 31 (M = 24.37 years). 20 of these were university students. Participants were recruited via postings on university-related Facebook groups and via email messages and printed posters on campus. The financial compensation varied between 9.40 and 10.40 euros. For the English version of the experiment, we tested 26 monolingual native speakers of American English, including 12 males and 14 females who were aged between 19 and 24 years old (M = 20.65 years). All were either undergraduates (n = 20) or graduate students (n = 6) in a Midwestern university, and were recruited via email messages. All undergraduates were enrolled in a first-year language class, and they were compensated for their time with extra points toward their final course grade, which corresponded to the amount of the financial gain earned in the experiment (varied between 8.80 and 11.40 dollars).

2.2 Materials

Before starting the experiment, participants had to read a context presenting them with an enacted betting scenario of a TV show and ensured all experimental stimuli were considered as part of a context rather than in isolation. Moreover, the context made very clear what the domain under the discussion was – the five contestants – and that all of them are relevant for the interpretation of the stimuli. The concrete task was for participants to judge bets as won or lost; judging a bet as "won" corresponded to accepting a target sentence. On each trial, participants saw experimental materials presented on slips with two sides. The front side of the slip included a bet concerning the happenings in the show, which they had to evaluate. The bet appeared in the form of a sentence containing an embedded question under *to surprise* or *to agree*, and it also contained a monologue/dialogue that expressed the beliefs of the attitude holder in question. In the case of *surprise*, the back side displayed a table summarizing what actually happened, i.e., the facts in the world. For the verb *agree*, this was not the case since there is no objective factual base against which to measure the attitude holders' subjective agreement. To create our experimental materials, we manipulated two factors. First, we manipulated READING, tested by changing the contents of the attitude holder's statement and the reported actual facts in the world. The readings tested depended on the embedding verb and appear in Table 1 & 2. As the semantic analysis of *surprise* is controversial (Roelofsen et al. 2019), we tested the same readings in both languages for this verb (Table 1). For *agree*, we tested partly different readings in English and German, since we had no reason to believe that there could be cross-linguistic differences. However, as *agree+Q* involves two AHs, there are a number of possible belief configurations that we wanted to cover in the experiments. Second, we tested the factor ROLE: In Role 1, participants had to redeem bets and would profit from bets that are won. In Role 2, they had to review betting slips and decide whether to pay out a reward. In this role, one profits from lost bets. To harness in the biases of the two roles, fees for redeeming bets, and fines for not paying out rewards for won bets were part of the rules. Thus, participants had a real financial incentive for answering correctly.

Our experiment had four crucial features: a) Correctness of the answers was evaluated post-experiment, such that no training artefact could emerge; b) the roles induced a different financial bias to judge bets as lost/won; c) participants had to reckon with negative financial consequences in case the bet was incorrectly judged as won/lost, thereby boosting reliability and robustness as a design feature; d) the use of direct financial incentive is known to increase effort (Camerer & Hogarth 1999). The linking hypothesis between participants' responses and readings is based on utility maximization in simple decision problems. Expected utility is measured in terms of direct financial payoff. The facts in the world and the financial gains/losses were correlated with the readings we tested (Table 1 & 2) such that the payoff, illustrated in Table 3, emerges for a person who decides to redeem/pay out a bet. Expected utility was calculated based on situation and on the probabilities of the three readings.

Table 3

	Role 1			Role 2		
Situation Reading	WE$_{\text{distr.}}$/ CA+	WE$_{\text{nondistr.}}$/ CASU	SE/CA/ IA	WE$_{\text{distr.}}$/ CA+	WE$_{\text{nondistr.}}$/ CASU	SE/CA/IA
WE$_{\text{distributive}}$ /CA	20	-10	-10	10	-20	-20
WE$_{\text{nondistrib}}$ /CASU	20	20	-10	10	10	-20
SE/CA-/IA	20	20	20	10	10	10

2.3 Analysis

We created two Bayesian statistical models for the experimental data. The first model, which we call the standard model, assumed that there is a fixed value of subjective probability for the three readings in the entire population. The second model, the variable-value model, considers the possibility that different persons have different probabilities for the different readings. We ran the models using a Hamiltonian Monte Carlo simulation with the no-U-turn sampling (NUTS) algorithm (Hoffman & Gelman, 2014) and performed model comparison using the Bayes factor (BF) and the bridgesampling package (Gronau, Singmann, & Wagenmakers, 2020). For the verb *surprise* the best model was the variable value-model, see Figure 1 for the posteriors. In both languages, the SE reading is the dominant interpretation, but there is also a group of people that has a non-distributive weak exhaustive interpretation of questions embedded under this verb. For German *agree*, the best model was the standard model with the CASU reading as the dominant interpretation. For English *agree*, the CA reading was dominant followed at a big distance by the CASU interpretation. The best model in this case was the variable-value model. Figure 2 shows the posterior of the best model for each language. Note that for German, we tested a negative baseline reading instead of the CA reading.

2.4 Conclusion

Our findings align with prior empirical evidence. They suggest reliability and robustness for the previously found readings. For *surprise*, the weakest reading (SE) is indeed dominant, cf. Cremers & Chemla (2017). For *agree*, we found that it is not required for the AHs to have the same opinions on the entire answer space, cf. Chemala & George (2016). Future research should take into account more the cross-linguistic landscape and investigate the effect of varying gains and losses in terms of prospect theory (Kahneman & Tversky 1979).

Figure 1

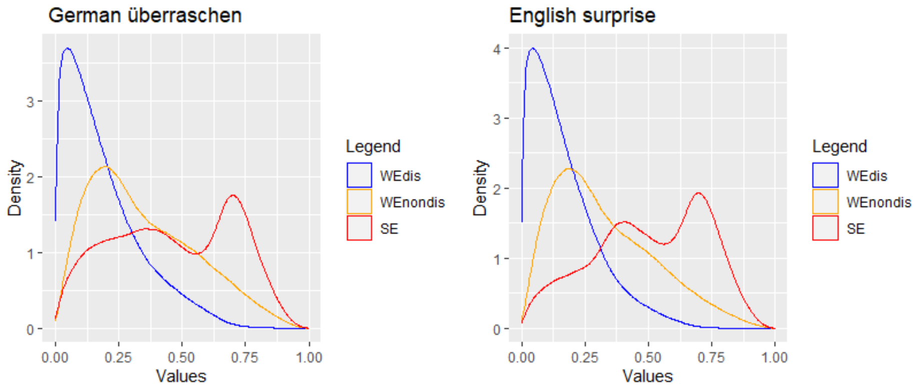
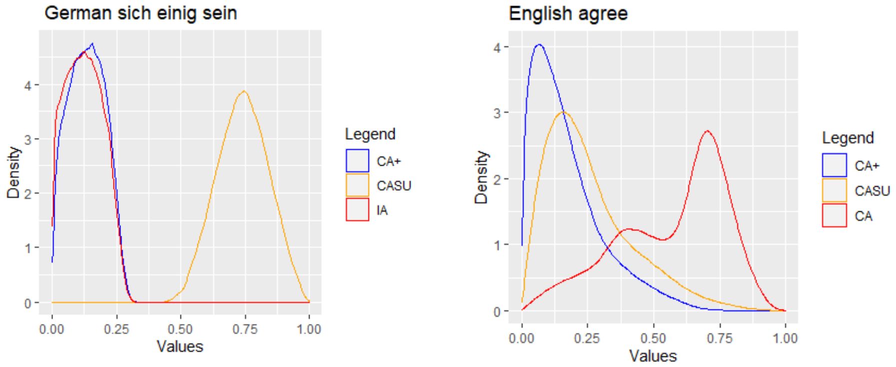


Figure 2



References

Beck, S. & Rullmann, H. (1999). A flexible approach to exhaustivity in questions. *Natural Language Semantics 7,* pp. 249-298.

Berman, S. (1991). On the semantics and logical form of Wh-clauses. PhD thesis. Amherst, MA: University of Massachusetts.

Camerer, C. F. & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty 19(1),* pp. 7-42.

Cremers, A. & Chemla, E. (2017). Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Natural Language Semantics 25,* pp. 223-236.

Chemla, E. & George, B. R. (2016). Can we agree about 'agree'? *Review of Philosophy and Psychology 7(1),* pp. 243-264.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7(4)*, pp. 457-472.

Gronau, Q. F., H. Singmann, and E.J. Wagenmakers (2020). "bridgesampling: An R Package for Estimating Normalizing Constants". In: Journal of Statistical Software 92.10, pp. 1–29.

Kahneman, D. & Tverskly, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47(2),* pp. 263-291.

Lahiri, U. (2002). Questions and answers in embedded contexts. Oxford: Oxford University Press.

Klinedinst, N. & Rothschild, D. (2011). Exhaustivity in questions with non-factives. *Semantics and Pragmatics 4(2)*, pp.1-23.

Kratzer, A. (2006). *Exhaustive Questions*. Linguistics Colloquium, MIT

Roelofsen et al. 2019. The *whether puzzle. In v. Heusinger et al. (eds.). *Questions in Discourse*, 172–197. Leiden: Brill.