Assessing classification reliability of conditionals in discourse

Reuneker, A.

Argumentation
https://doi.org/10.1007/s10503-023-09614-9

ORIGINAL RESEARCH

Assessing Classification Reliability of Conditionals in Discourse

Alex Reuneker[1]

Received: 19 November 2022 / Accepted: 16 March 2023

Abstract

Conditional constructions (*if–then*) enable us to express our thoughts about possible states of the world, and they form an important ingredient for our reasoning and argumentative capabilities. Different types and argumentative uses have been distinguished in the literature, but their applicability to actual language use is rarely evaluated. This paper focuses on the reliability of applying classifications of connections between antecedents and consequents of conditionals to discourse, and three issues are identified. First, different accounts produce incompatible results when applied to language data. Second, a discrepancy between theory and data was observed in previous studies, which sometimes discard existing classifications for being detached from actual language use. Finally, language users construct various cognitive relations between clauses of conditionals without being able to rely on overt linguistic features, which poses problems for the annotation of conditionals in argumentation and discourse. This paper addresses these issues by means of comparing theoretical types and actual uses of conditionals, by inspecting the dispersion of types in natural-language corpora, and by conducting an experiment in which the inter-rater reliability of classifications was assessed. The results show that the reliability of classifications of conditionals when applied to language data is low. With respect to the aforementioned issues, different classifications produced incompatible results, a discrepancy between theory and data was indeed observed, and low reliability scores indicated a largely interpretative nature of types of conditionals. Given these results, suggestions for the enhancement of reliability in corpus studies of conditionals and beyond are provided to enhance future classification design.

Keywords Conditionals · Inter-rater reliability · Corpus · Classification

✉ Alex Reuneker
a.reuneker@hum.leidenuniv.nl

[1] Leiden University Centre for Linguistics (LUCL), Leiden University, Postbus 9515, 2300 RA Leiden, The Netherlands

Published online: 06 April 2023

1 Introduction

Conditional constructions enable us to express our thoughts about possible states of the world, and they form an important ingredient of our cognition and argumentative capabilities. According to Evans and Over (2004, p.1), conditionals form 'an essential part of human reasoning and decision making'. Edgington (2022, p. 188) argues that they are 'essential to practical reasoning about what to do, as well as to much reasoning about what is the case'. Conditionals are involved, as Hartmann and Hahn (2020, p. 981) mention, in 'every aspect of our thinking, from the mundane and everyday such as "if you eat too much cheese, you will have nightmares" to the most fundamental concerns as in "if global warming isn't halted, sea levels will rise dramatically"'. As is known from the linguistic literature on conditionals, however, defining what exactly constitutes a conditional is "extremely difficult" (Declerck & Reed 2001, p. 8), or even 'impossible' (Wierzbicka 1997, p. 16). Even without a definition, different types and argumentative uses of conditionals have been distinguished (for an overview, see Reuneker 2022a). Dancygier and Sweetser (2005), for instance, argue for a distinction between predictive conditionals, in which antecedents and consequents are causally related, as in (1), and non-predictive conditionals, such as inferential and speech-act conditionals, in which the clauses present an inference chain from argument to conclusion, or a contextualization and a speech-act, as in (2) and (3) respectively.

(1) If you mow the lawn, I'll give you ten dollars.
(2) If he typed her thesis, he loves her.
(3) If you need help, my name is Ann.

Athanasiadou and Dirven (1997a), on the other hand, distinguish between course-of-event, hypothetical and pragmatic conditionals, as in (4) to (6).

(4) If there is a drought like this year, the eggs remain dormant.
(5) If the weather is fine, we'll go for a swim.
(6) If you are thirsty, there's beer in the fridge.

Whereas some types are identical in various accounts, such as the speech-act or pragmatic type in (3) and (6), it remains unclear how other types are related, and how they impact the reasoning tasks and argumentative maneuvers for which they are used. Before answering such questions, however, another issue should be addressed.

Various corpus studies on conditionals criticize existing top-down (deductive) classifications, such as those presented above, for being too detached from actual language use (see e.g., Carter-Thomas and Rowley-Jolivet 2008), or for being too dependent on contextual interpretation (see e.g., Ferguson 2001). In turn, this criticism has led to several smaller-scale bottom-up (inductive) classifications, which better suit the data under investigation, but prohibit more general conclusions and replication. Claims such as the one by Dancygier and Sweetser (2005, p. 137), in

which they remark that frequencies of different types of conditionals 'vary radically depending on the subject matter and the speaker's or author's goals', can only be tested properly if there is a reliable way of identifying such types in different datasets. As a result of projecting theoretically motivated categories onto language data, and applying categories constructed based on one language or text genre onto another, corpus research risks ongoing formation of new categories and ever shifting boundaries. Whereas this issue is explored extensively in some areas of discourse analysis and argumentation research, such as framing analysis (see e.g., van Gorp 2007), this is not the case for conditionals.

Three issues are identified in this paper. First, as exemplified above, different accounts produce incompatible results when applied to actual language data. The second issue is that of a discrepancy between theory and data—discourse studies sometimes discard existing classifications for being "too detached" from actual discourse (e.g., Carter-Thomas and Rowley-Jolivet 2008). Finally, while language users construct various cognitive relations between the clauses of conditionals to argue, for instance, from cause to effect, or argument to conclusion, they do so without being able to rely on overt linguistic features (see Reuneker 2022a, b), which poses problems for the annotation of conditionals in argumentation and discourse. Reliability is a prerequisite for the demonstration of the validity of a classification scheme, i.e., showing 'that the coding scheme captures the "truth" of the phenomenon being studied' (Artstein and Poesio 2008, p. 557), including issues such as the cognitive plausibility of types distinguished within a classification (for a recent discussion on descriptive adequacy and cognitive plausibility, see e.g., Scholman et al. 2022). These issues warrant a critical assessment of the applicability of classifications of conditionals to natural language data, and issues of reliability in any discourse-oriented study aimed at the cognitive mechanisms underlying interpretation and argumentation. The aim of this paper is therefore to evaluate the application of classifications of conditionals to natural language corpora. To do so, an experiment was conducted to answer the following research question: to what extent can existing classifications of natural-language conditionals be applied reliably to corpus data?

In Sect. 2, three classifications of conditionals that are used in the experiment central in this contribution will be discussed. In Sect. 3, the data and method are discussed, and in Sect. 4, the results are presented. In Sect. 5, an answer to the research question is provided, conclusions with regard to the issues raised above are drawn, and the ramifications for future research are discussed.

2 Classifications of Conditionals

Conditionals have been studied 'since Aristotle' (Dancygier 1998, p. 1) and it is not surprising that several disciplines, ranging from philosophy (see Bennett 2003) to psychology (see Oaksford & Chater 2010), have investigated the subject in its many facets. Within the study of language and argumentation, classifications by Quirk et al. (1985), Athanasiadou and Dirven (1997a) and Dancygier and Sweetser (2005)

have been influential. Central to these classifications is the relation between antecedent (*if*) and consequent (main clause).[1]

As many authors note, the applicability of classifications is challenging and an important measure of its validity (see e.g., Artstein and Poesio 2008, p. 557; Bolognesi et al. 2017, p. 1993). The classifications under investigation in this contribution will be discussed below, in order to explain the classification schemes used in the experiment reported in the next sections.[2]

2.1 Direct and Indirect Conditionals

Quirk et al. (1985) distinguish between two main categories of conditionals: *direct* and *indirect conditionals*, comparable to *semantic* and *pragmatic* discourse coherence relations (cf. Sanders et al. 1992). Conditionals expressing a direct condition indicate that 'the truth of the proposition in the matrix clause is a consequence of the fulfilment of the condition in the conditional clause' (Quirk et al. 1985, p. 1088). It has two subtypes: *open* and *hypothetical conditions*. The conditional clause in (7) is *open*, as it is neutral with regards to the fulfilment of the condition.[3]

(7) If you put the baby down, she'll scream.

The second subtype of direct conditionals is called *hypothetical*, because it expresses negative epistemic stance towards the fulfilment of the condition in the antecedent, as in (8).

(8) If you had listened to me, you wouldn't have made so many mistakes.

The *if*-clause is marked for modality, which, based on the connection between antecedent and consequent, is carried over to the main clause.

Indirect conditionals express a condition that is not directly related to the situation in the main clause. The subordinate *if*-clause can express, for instance, a metalinguistic comment, as in (9), or the condition under which the speech act in the main clause is uttered, as in (10).

(9) His style is florid, if that's the right word. […].
(10) If you're going my way, I need a lift back.

[1] The 'connectedness' of antecedents and consequents is considered here to be one of two conventional meaning aspects of conditionals, next to unassertiveness. The specific connections, which form the focus of this contribution, are considered conversational implicatures (cf. Grice 1989). Details of this analysis can be found in Reuneker (2022a, Chapter 2).

[2] As these classifications strive to be exhaustive, a full discussion is outside the scope of this contribution.

[3] Unless otherwise noted, the examples in this section are taken from the respective accounts.

The final type discussed by Quirk et al. (1985, pp. 1094–1095) is the *rhetorical conditional.* It has the appearance of open conditionals, as in (7), but makes 'a strong assertion'. In case of (11), the absurdity of the proposition in the consequent is projected onto the antecedent, in turn rendering it false.[4]

(11) If they're Irish, I'm the Pope. (Since I'm obviously not the Pope, they're certainly not Irish.)

2.2 Hypothetical, Course-of-Event, and Pragmatic Conditionals

Athanasiadou and Dirven (1997a, b) distinguish between *hypothetical*, *course of event*, and *pragmatic conditionals*. Their account is the result of interpreting the distribution of features of conditionals in several corpora (see Athanasiadou & Dirven 1997a, p. 63). Prototype theory is used to distinguish between types, meaning that features of conditionals are not seen as necessary and sufficient conditions, but as more or less prototypical features (cf. Rosch 1978). The main feature used is *hypotheticality*, which is prototypical for only one specific (albeit highly frequent) type of conditional.

Hypothetical conditionals operate in what Athanasiadou and Dirven (1997a, p. 62) call a 'non-actual frame' and are non-assertive with respect to the antecedent and the consequent. Hypothetical conditionals express two different events, and it is the events—not their relation—that are marked as hypothetical. The relation is prototypically *causal,* as in (12).

(12) If there is no water in your radiator, your engine will overheat immediately.

In addition to a cause, the antecedent can also present a *condition*, licensing a second subtype, which is still causally related to the consequent, but less strongly so than the former subtype, as in (13).

(13) If the allowance is more favourable to a widow than the retirement pension, she will be paid that allowance.

The third hypothetical subtype is *supposition*, as exemplified in (14), in which the consequent expresses a reaction to a 'supposed state of affairs' (Athanasiadou and Dirven 1997a, b, p. 66). Here, even less causality is involved; there is merely a possible 'resultative action'.

(14) If I go bald I'll shoot myself.

[4] In contrast to what Gabrielatos (2010, p. 156) argues, rhetorical conditionals are not considered a sub- or special type of direct conditions here, because Quirk et al. (1985, p. 1094) argue that they are (strongly) assertive; a characteristic opposed to the non-assertiveness of open conditions.

The second main type is the *course-of-event conditional*, which asserts a relation of co-occurrence between two situations (i.e., a *whenever* relation), as in (15).

(15) If there is a drought like this year, the eggs remain dormant.

In contrast to hypothetical conditionals, the speaker commits himself to the 'actual, frequent or general realization of the two situations.' This category is divided into three subcategories: *descriptive*, *inferential* and *instructive course-of-event conditionals*, exemplified in (15) to (17) respectively.

(16) He looked at his watch; if the soldier was coming, it was nearly time.
(17) You should call a doctor if there is a fever.

In (16) the consequent holds a conclusion based on the antecedent. In (17), the consequent holds an instruction dependent on the assumption in the antecedent. Common to all course-of-event conditionals is their generic or iterative nature.

The last main type, the *pragmatic conditional*, presents an indirect relation between antecedent and consequent; i.e., the antecedent expresses a 'meta-pragmatic signal'. This low level of dependency between antecedent and consequent has consequences for their syntactic integration, as was noted earlier by Quirk et al. (1985) for *indirect conditions*. Pragmatic conditionals are subdivided into *conversational conditionals*, operating on the level of speech acts, as in Austin's (1961, p. 210) example in (18), and *logical conditionals,* involving 'analytic reasoning processes', as in (19).

(18) There are biscuits on the sideboard if you want them.
(19) If she's divorced, then she's been married before.

As logical conditionals communicate how one moves from an argument to a conclusion, they are frequently accompanied by linguistic features such as *then* and modal verbs like *must* to emphasize 'the act of reasoning.'

2.3 Predictive and Non-predictive Conditionals

Dancygier and Sweetser (2005) characterize the relation between antecedents and consequents of conditionals in terms of different domains. The main distinction is that between *predictive* and *non-predictive* conditionals. Predictive conditionals are used for predictive reasoning, as in (20).

(20) If it rains, the match will be cancelled.

Because the prediction is based on 'real-world' enablement or causality (cf. Van der Auwera 1986), these conditionals are also called *content conditionals.* Dancygier and Sweetser (2005, p. 27) consider *content conditionals* prototypes for the other types (see below), in which causality has been pragmatically extended to the (epistemic) domain of reasoning, and of speech-acts. Predictive conditionals

have a *backshifted* verb in the antecedent, marking a time that 'is earlier than the time actually referred to' in order to form the background to the prediction in the consequent (Dancygier 1998, p. 37). For (20) for instance, Dancygier (1998, p. 38) argues that 'it rains' indicates the present but refers to the future (which would require *will* in the antecedent). The degree of backshift correlates with the epistemic distance marked by the speaker, from neutral stance, as in (20), to hypothetical stance as in (21).

(21) If it rained, the match would be cancelled.

Next to backshifted conditionals, Dancygier and Sweetser (2005, p.95) consider *generic conditionals*, as in (22), to be a subtype of content conditionals, expressing a *whenever*-relation.

(22) He gets angry if I leave the house.

The second type, and the first non-predictive type, is the *epistemic conditional,* which is one step removed from real-world causality. Instead of expressing a content-level prediction, this type expresses causality at the level of reasoning processes: the (hypothetical) knowledge of the truth of the antecedent is a sufficient condition for drawing the conclusion in the consequent, as in (23).

(23) If he typed her thesis, he loves her.

(My knowledge that the typing happened is a precondition for my conclusion about the loving.)

The function of the antecedent of epistemic conditionals is 'simply to give background to the addressee, by invoking the relevant parts of the cognitive context which brought about this conclusion' (Dancygier & Sweetser 2005, p. 117). Sweetser (1990, p. 123) characterizes epistemic conditionals as expressing 'reversed causality'; the antecedent functions therefore as an argument for the conclusion in the consequent and can be rephrased as 'It is known that P, therefore I can/may conclude Q'. The type is accompanied by a high-frequency of *then* and non-deontic use of modals like *must* to signal reasoning processes.

Speech-act conditionals are yet another step away from real-world causality, as the speech-act in the consequent is conditional on the fulfilment of the felicity condition in the antecedent. Sweetser provides the famous example by Austin, as presented in (18) above, in which the maxim of relation is invoked; only in the case of the hearer being hungry the offering of biscuits is relevant. As a general characterization, it can be said that the consequents of speech-act conditionals involve felicity conditions rather than truth values and they include questions, commands, requests and assertions as consequents. As with epistemic conditionals, the connection between antecedent and consequent in speech-act conditionals is indirect, as the antecedent 'merely' contextualizes the uttering of the consequent. As such, this type does normally not occur with distanced verb forms, and they do not license

conditional perfection, i.e., inferring 'if not P, not Q' from 'if P, not Q' (cf. Geis and Zwicky 1971). It must however be distinguished from Athanasiadou and Dirven's *pragmatic conditionals*, which include a much wider range of types.

Finally, Dancygier and Sweetser (2005) discuss the metalinguistic conditional, as in (24).

(24) That seems to be the gist of my thinking, if acting solely on impulse can be called thinking.

Such conditionals differ from speech-act conditionals in relating to the linguistic form of the speech act, not to its illocutionary force.

2.4 Classifications Compared

The classifications discussed all strive for an exhaustive set of types. Although Athanasiadou and Dirven present their classification in terms of prototype theory, as does Dancygier (1998), in their results every conditional sentence is classed as belonging to one type, which shows that the authors implicitly strive for mutually exclusive types. As Sweetser (1990, pp. 124–125) argues, 'a given example may be ambiguous between interpretations in two different domains, […], but no one interpretation of an *if–then* sentence […] simultaneously expresses conditionality in more than one domain'. However, prototype categories can have 'fuzzy boundaries' (cf. Taylor 2003, p. 51) and, in this case, a conditional sentence may simultaneously display features of more than one type.

With respect to the first issue raised in the introduction, it is insightful to see where the three classifications overlap and where they differ. A small sample examples from the accounts discussed was classified according to the main types of the three classifications discussed (Table 1).

In this table, it can be seen that for cases 4 and 5, the classifications consistently characterize these conditionals as *indirect*, *pragmatic* or *speech-act*. However, this does not apply to all cases, as can be seen in case 1, for which Quirk et al. (1985, p. 1094) have the designated type *rhetorical conditional*, while for Athanasiadou and Dirven the example is belongs to a subtype of *pragmatic conditionals*. As Quirk et al. (1985) place rhetorical conditionals outside their *direct–indirect* distinction, and pragmatic conditionals would fall inside their *indirect* class, this adheres to an inconsistency between classifications. Dancygier and Sweetser (2005) do not analyze rhetorical conditionals as a separate type, but the example satisfies the criteria of epistemic conditionals, as the falsity of the antecedent licenses the conclusion in the consequent. Epistemic conditionals are a subtype of non-predictive conditionals, however, while they are direct conditions in Quirk et al. (1985, p. 1091) and a subtype of course-of-event or pragmatic conditionals in Athanasiadou and Dirven's classification. This brief comparison shows that classifications of the same phenomenon may produce incompatible results.

Table 1 Conditionals classified according to three classifications

	Conditional sentence	Quirk et al. (1985)	Athanasiadou and Dirven (1997a)	Dancygier and Sweetser (2005)
1	If that's art, then I'm an artist too!	*Rhetorical*	*Pragmatic*	*Epistemic*
2	If Oscar is selected to study medicine, then a golden future awaits him	*Direct*	*Hypothetical*	*Predictive*
3	If life is a candle, people are the moths flying towards it	*Indirect*	*Pragmatic*	*Metalinguistic*
4	So if I understand correctly, the singer of REM did not leave the band?	*Indirect*	*Pragmatic*	*Speech-act*
5	So: if you're interested and you don't have any plans yet, the Dutch Philharmonic Orchestra plays Tchaikovsky tonight	*Indirect*	*Pragmatic*	*Speech-act*
6	If there wouldn't be coffee, the guests wouldn't stay this long	*Direct*	*Hypothetical*	*Predictive*
7	And every time you have worked-out, your energy level has dropped	*Direct*	*Course-of-event*	*Predictive*
8	If the O is in the name of the month again, it'll get cold quickly	*Direct*	*Course-of-event*	*Epistemic*

2.5 Conclusion

In this section, three classifications were discussed. Although the classifications are targeted at the same phenomenon, they do not produce identical results. Discrepancies between classifications are, in themselves, not problematic, however; as long as classifications are viewed as artificial constructs rather than reflections of natural systems (cf. Sandri 1969, pp. 86–87), different perspectives and organizations can co-exist. This shifts the common question 'which classification is right?', to 'which classification is able to explain the data best and most efficiently?' Preliminary to those questions, however, are questions of validity and reliability, which will be discussed in the remainder of this contribution.

3 Data and Method

While the classifications discussed above are aimed at offering a descriptive account of types of conditionals found in natural language, their reliable applicability to language data is not discussed in detail. As Spooren and Degand (2010, p.242) remark, 'there is presently no tradition in the field of corpus-based discourse studies to report agreement measures'. Athanasiadou and Dirven (1997a), for instance, provide frequencies of their attested types, but do not mention how these results were obtained and whether or not the annotations were checked for consistency. Dancygier and Sweetser (2005) use examples from corpora, but no reliability measures are provided.

Reliability is of major importance to the study of conditionals, as the assignment of instances to classes of conditionals is, inevitably, based (at least partly) on interpretation. Reuneker (2022a), for instance, considers the relations discussed in the previous section as conversational implicatures (cf. Grice 1989), which are, per definition, context-dependent to some degree. The notion of reliability is therefore vital for the assessment of the extent to which classification results are 'independent of the measuring event, instrument or person' (Kaplan and Goldsen 1965, p. 83). This section describes the method with which the reliability of classification of conditionals in corpus data is evaluated.

3.1 Evaluation of Reliability

In this study, the concept of reliability was investigated, as high reliability suggests findings that are unaffected by individual differences in raters. Reliability here is understood as the combination of *stability* (do raters' judgments remain constant over time) and *replicability* (can judgments be reproduced among raters). As such, it differs from measures of *validity*, which represent the 'the extent to which [both] raters classify subjects into their true category' (Gwet 2014, p. 314).

To evaluate the applicability of classifications of conditionals, an experiment was carried out in which a group of trained participants classified a set of conditionals

from the CONDIV-corpus (Deygers et al. 2000) of written Dutch and the CGN corpus of spoken Dutch (Oostdijk 2000). While some corpus studies of conditionals, such as the large-scale annotation in the Penn Discourse Tree Bank 2.0 project (Prasad et al. 2007, p. 1), report agreement, they do so in percentages, which are not corrected for chance-agreement. As is shown in the literature on reliability (see e.g., Gwet 2014), a skewed distribution of types and subtypes may cause overestimating reliability scores. When one type, such as the hypothetical conditional in the Penn Discourse Tree Bank 2.0 project, comprises a large majority of cases, the chance of a labeling a hypothetical condition as such is higher than labeling, for instance, the less frequent type of relevance condition correctly. In other words, if a rater would not look at the actual example and/or instructions and would class it as hypothetical, the chance of this being correct would be significantly higher than random assignment. Chance agreement needs to be corrected for, which is done in this study using Krippendorf's Alpha (Hayes and Krippendorff 2007; see Sect. 4.2). The design, participants, materials, and procedure of the current experiment are described in the following sections.

3.2 Design

The experiment followed a within-participants design to control for possible effects of individual differences in linguistic knowledge and understanding of the materials. All participants classified 33 items according to the three classifications discussed in Sect. 2. To control for memory and practice effects, the order of blocks (i.e., part of the annotation session in which one specific classification is applied) was counterbalanced using a latin-square design.

3.3 Participants

The participants were 27 students of Linguistics at Leiden University (22 female, 5 male) with an average age of 22.7 years ($sd = 5.13$). All participants (*raters* in following sections) were native speakers of Dutch. The students participated for course credit in a course on conditionals.

3.4 Materials

The items were Dutch conditional sentences and consisted of 3 examples (labeled as such) to familiarize raters with the task, 14 items from the written corpus, and 9 items from the spoken corpus. Furthermore, there were 8 control items, i.e., variations of textbook examples to enable selection of raters (see Sect. 4.1), 2 test–retest items to be able to evaluate intra-rater reliability, i.e., the stability of annotation over time per rater.[5] The sentences were selected from a random sample of sentences containing the Dutch

[5] The complete set of items, as well as scripts and data, can be found online (see Reuneker 2022b).

conditional conjunction *als* 'if'. The sample was cleared of non-conditional uses of the conjunction, such as the comparative use of *als* 'like' and temporal uses of *als* 'when'.

All items included one sentence preceding and one sentence following the conditional sentence. The conditional sentence itself was presented in bold. Examples of written and spoken corpus items including context are presented in (25) and (26) respectively.

(25) Veel mensen doen dingen waarvan ze best weten dat het niet mag. **Maar als het niet wordt bestraft, gaat zoiets wennen en gaat het steeds een stukje verder.** Daar gaan we nu een stokje voor steken. (NRC, nie_sp1.txt)

Many people do things while knowing they're not allowed to. ***But if it isn't punished, people will get used to it and it will continue to progress.*** *That is what we are going to prevent.*

(26) oh. ja ik vind dat wel heel interessant en ik **als ik dat zo zie dan denk ik ook dat die man 't heeft gedaan.** ook als je dat hoort maar van de andere kant ja. (fna000458__166)

oh. Yeah I find that very interesting and I ***if I see it like that then I also think that man has done it.*** *Also if you hear that but on the other side yes.*

3.5 Procedure

All raters were students taking a course on conditionals, in which, for each classification discussed in Sect. 2, the original article or book section was distributed as part of the course materials. Raters were asked to read the text and classify a set of conditionals accordingly prior to the class meeting in which the classification was discussed. Both the examples provided by the authors and real usage data were used as training material. A week before the experiment, raters were presented with an overview of the classifications, including criteria for each type (see Reuneker 2022b), in order to enable the participants to evaluate their understanding of the source texts and familiarize themselves with the instructions for the classification task. At the start of the experiment, subjects were presented with instructions on paper (see Reuneker 2022b). The items were implemented in Qualtrics.

The raters first classified three examples to familiarize themselves with the task at hand. After these examples, each participant classified 23 corpus items, 8 control items and 2 test–retest items. Per item, participants selected a type of conditional, indicated their confidence on a 5-point Likert scale and optionally included a remark. After classifying all items, raters were asked to provide their name, age, first language, and study program. In total, all raters classified 102 sentences. The time spent by raters on the total experiment was on average 49.85 min ($sd = 11.23$).

Table 2 Accuracy for control items per classification (before selection, n = 27)

Classification	Mean accuracy	Sd
Quirk et al. (1985)	0.84	0.12
Athanasiadou and Dirven (1997a)	0.81	0.17
Dancygier and Sweetser (2005)	0.68	0.16

Table 3 Accuracy for control items per classification (after selection, n = 18)

Classification	Mean accuracy	Sd
Quirk et al. (1985)	0.89	0.06
Athanasiadou and Dirven (1997a)	0.83	0.15
Dancygier and Sweetser (2005)	0.75	0.13

Table 4 Agreement per classification for control and corpus items

Classification	Control (main)	Control (sub)	Corpus (main)	Corpus (sub)
Quirk et al. (1985)	0.87	0.69	0.53	0.41
Athanasiadou and Dirven (1997a)	0.59	0.45	0.31	0.29
Dancygier and Sweetser (2005)	0.55	0.56	0.32	0.28

4 Results

The experimental data were analyzed as follows. First, the validity of ratings on the control items was computed for each rater. Second, agreement on corpus items was calculated for each classification. Third, a distribution of pairwise agreement scores was generated to enable more detailed comparison of classifications. Fourth, the stability of class assignments was evaluated, and finally, agreement scores were compared to self-indicated values of confidence to test for a possible correlation. Each of these steps is described in the sections below respectively.[6]

4.1 Selection of Raters and Evaluation of Control Items

Eight control items (see Sect. 3.4) were randomly presented in each of the three trials. These items were constructed in adherence to the criteria presented in the classifications discussed, resulting in idealized examples. No authentic examples by the respective authors were included to avoid memory effects. As the goal of this study was to measure the reliability of existing classifications when applied to actual language data, it was found necessary to select only those raters who were able to correctly classify idealized examples.

[6] For a pair-wise comparison on item level, see Reuneker (2022a, pp. 168–172).

Measuring chance-corrected agreement between raters assumes the lack of a so-called 'gold standard'; the correct classifications of the items in question (see e.g., Gwet 2014, p. 19). For the selection procedure described here, however, a gold standard was available. Therefore, not reliability, but *accuracy* was calculated for each rater. Accuracy was calculated by dividing the number of correct answers by the total number of classifications made. The results are presented in Table 2 below.

Instead of using an arbitrary cut-off point, negative deviation from the mean accuracy was used for selection of raters. If a rater's accuracy score was more than one standard deviation from the mean (a z-score of -1 or less), this was taken to signal inadequate understanding of the task at hand. As can be seen in Table 3, nine subjects were excluded from further analysis, resulting in higher accuracy scores and lower deviations (i.e., smaller differences between raters).

A repeated-measures ANOVA on the data from Table 4 ($F(2,36) = 7.58$, $p = 0.002$) showed that classification had a significant effect on accuracy within each subject. A post-hoc test using Bonferroni correction confirms this difference to be significant ($p < 0.001$) for comparisons with Dancygier and Sweetser's classification, meaning that their accuracy (0.75) was significantly lower than that of Quirk et al. (0.89) and Athanasiadou and Dirven (0.83). In other words, raters had more difficulty classifying idealized examples of conditionals using Dancygier and Sweetser's classification.

4.2 Inter-Rater Reliability on Corpus Data

In contrast to the accuracy of control items in the previous section, no gold standard was available for the corpus data, and agreement was calculated. The use of raw percentages of agreement is heavily debated (see e.g., Banerjee 1999), as they do not take into account that agreement can be reached by chance (cf. Cohen 1960). Therefore Krippendorf's Alpha (Hayes and Krippendorff 2007) was calculated, which corrects for chance agreement and is comparable in interpretation to the more familiar Cohen's Kappa (for details, see Reuneker 2022a, pp. 194–198). The results are presented in Table 4 for both main types and subtypes of conditionals (see Sect. 2).[7]

This table shows that the agreement on corpus items is low to moderate [8]overall (see Landis and Koch 1977), as even the highest score on corpus items signals

[7] To compare measures of reliability, Gwet's AC1 values, which are chance corrected in a different way (see Gwet 2014, pp. 104–105, 118–119), were also calculated for the control and corpus items on main types. The resulting figures show the same trends; for control items using Quirk et al. (1985), Athanasiadou and Dirven (1997a), and Dancygier and Sweetser (2005), AC1 values were 0.85, 0.63, and 0.57 respectively, and for corpus items, the figures were 0.64, 0.35 and 0.39.

[8] The small and reversed difference between main level and sub-level test items for Dancygier and Sweetser can be explained by the small difference between the occurrence of four categories in the leveled results, and five categories in the non-leveled results. The reason that this account is not leveled to two categories (i.e., predictive, or *content* and non-predictive) is that studies using Dancygier and Sweetser's (2005) classification distinguish mainly between content, epistemic, speech-act and metalinguistic conditionals, not between predictive and non-predictive. This is not the case for Quirk et al. (1985), for which the biggest difference can be observed between the control and corpus items. This is affected by the fact that this classification has three main and eight sub-level categories. As this is a direct consequence of the theoretical foundations of a classification, it will not be controlled for in this study.

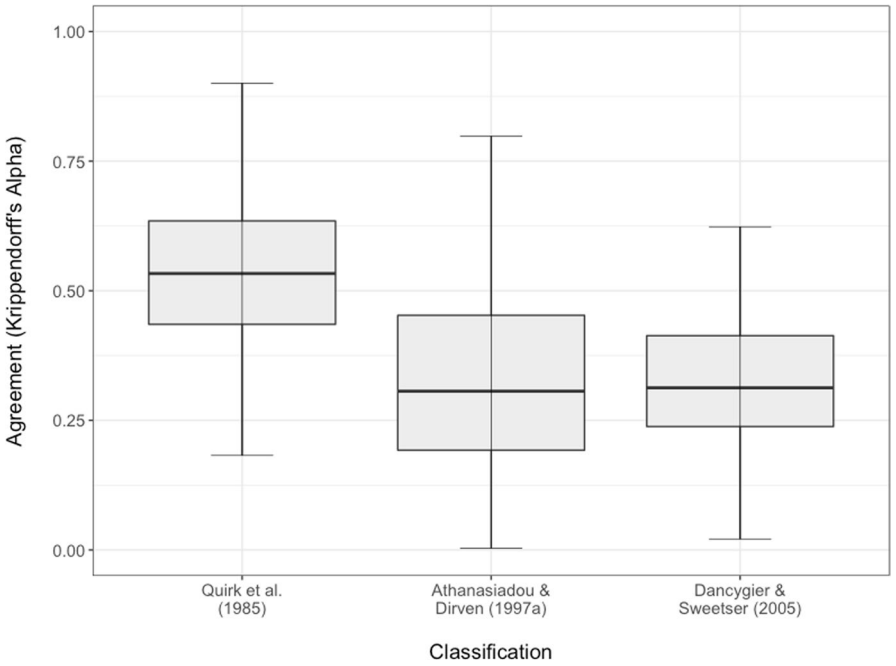

Fig. 1 Distribution of agreement coefficients per classification

only 'moderate agreement'. A comparison of agreement scores for control items and corpus items shows that raters were better in classifying idealized examples than in classifying attested conditionals; agreement on control items was higher (0.55–0.87), than on corpus items (0.31–0.53).

Both on control items and on corpus items, Quirk's classification results in substantially higher agreement scores than Athanasiadou and Dirven's and Dancygier and Sweetser's classifications. What can also be seen, is that, when subtypes are taken into account, reliability decreases, which is consistent with other observations in the literature (see e.g., Spooren and Degand 2010; Bolognesi et al. 2017, pp. 1993–1994). In the results presented and discussed below, only main types are considered.

4.3 Distributive Analysis of Agreement Coefficients

The results presented above provides only a general picture of the reliability of each classification. What is needed, is a way to statistically compare agreement scores between classification schemes. To allow for such a more detailed analysis, a novel approach to agreement analysis was devised. For each classification, a distribution of agreement coefficients was generated. For all combinations of raters, Krippendorff's Alpha was calculated, resulting in 153 coefficients per classification. These distributions allowed for a more detailed statistical analysis and comparison, because not

Table 5 Intra-rater reliability on test–retest pairs adapted from corpus items

Classification	Agreement (%)
Quirk et al. (1985)	91.7
Athanasiadou and Dirven (1997a)	77.8
Dancygier and Sweetser (2005)	91.7

only mean reliability, but also deviations were taken into account. The results are presented in Fig. 1 below.

A repeated-measures ANOVA showed that the independent variable *classification* had a significant effect on the dependent variable *agreement* ($F(2453) = 37.43$, $p < 0.001$). A post-hoc test using Bonferroni correction confirms this difference to be significant ($p < 0.001$) for comparisons between Quirk et al. (1985) on the one hand and Athanasiadou and Dirven (1997a) and Danycgier and Sweetser (2005) on the other. This means that the reliability of Quirk et al.'s classification (0.53) is significantly higher than that of Athanasiadou and Dirven (0.31) and Dancygier and Sweetser (0.32). Raters thus had more difficulty reliably classifying conditionals from corpora when using Athanasiadou and Dirven's and Dancygier and Sweetser's classifications.

4.4 Intra-Rater Reliability on Corpus Data

Whereas *inter*-rater reliability is concerned with the agreement between different raters, *intra-rater* reliability is concerned with the 'self-reproducibility' (Gwet 2014, pp. 6, 200) or *stability* of classifications—also called *test–retest reliability*. Intra-rater reliability is commonly seen as a weaker measurement of reliability than inter-rater reliability, because it only measures the degree to which classification results can be replicated by one rater, instead of by different raters. However, as, for instance, Verhagen and Mos (2016, p. 336) argue, the processing of linguistic material of an individual may vary between moments, which calls for the measurement of 'individual variation and its underlying dynamics'.

For each classification, one item from the spoken corpus and one from the written corpus was adapted to function as a test–retest pair. To rule out possible confounding variables, care was taken to apply changes only on the lexical-semantic level of the utterance. For a full inquiry into the stability of classifications, the calculation of intra-rater reliability should be chance corrected. As this study's focus is on inter-rater reliability, however, the number of test re-test items was limited to keep the task manageable for participants, and consequently, only percentages of intra-rater agreement were calculated. The results are shown in Table 5 below.

The percentages suggest that raters' judgments are stable and low inter-rater agreement scores are not the result of random assignment of items to classes. It should be noted, however, that Athanasiadou and Dirven's classification scores lower on intra-rater reliability, indicating less consistency within individual rater's annotations.

4.5 Agreement and Rater Confidence

In addition to classifying conditionals, raters also reported their confidence in the class chosen on a 5-point Likert scale (1 = very uncertain, 5 = very certain). A Pearson correlation coefficient was calculated to assess the correlation between agreement and confidence. There was a positive correlation between the two variables ($r(21) = 0.76$, $p < 0.001$), meaning that items that reached low agreement were found harder to classify by raters, in turn suggesting that raters were aware that certain items were harder to classify than others.

4.6 Conclusion

In this section, the results of an experiment comparing inter-rater agreement scores for each of the classifications discussed in Sect. 2 were presented. Raters achieve high accuracy scores for idealized examples of conditionals, whereas they cannot distinguish types of conditionals in corpus data reliably. The classification by Quirk et al. (1985) produced significantly higher agreement scores than the classifications by Athanasiadou and Dirven (1997a) and Dancygier and Sweetser (2005). An evaluation in term of intra-rater agreement showed raters' judgments to be stable, which suggests that the low inter-rater agreement scores are not the result of random assignment of items to classes. This is corroborated by the fact that raters could consistently identify conditionals they found hard to classify.

5 Conclusion and Discussion

This contribution started out with the question to what extent existing classifications of natural-language conditionals can be applied reliably to corpus data. The results suggest a predominantly negative answer to this question, as raters were not able to agree on the classification of conditionals in actual language data.

In the introductory section, three issues related to the question above were identified. First, different classifications produce incompatible results when applied to actual language data. The comparison of classifications indeed showed such incompatibilities, but it was also argued that they are not problematic per se when classifications are seen as artificial constructs rather than reflections of natural systems. This does raise questions concerning cognitive plausibility of the types distinguished, however. Shifting from a natural to an artificial view of classification directs evaluation from the question which classification is correct to which one is best able to explain the data. Second, several corpus studies on conditionals observed a discrepancy between theory and data. This study showed that all three classifications strive for exhaustiveness, yet a large number of conditionals from actual corpus data seem to resist the proposed types. This suggests there exists indeed a discrepancy between theory and data. Third, language users construct various cognitive relations between the clauses of conditionals to argue, for instance, from cause to effect, or argument to conclusion, but they do so without being able to rely on overt linguistic features.

The low reliability scores presented in this study indeed indicate such an interpretative nature of connections between antecedents and consequents of conditionals. In the following sections, ramifications and possible explanations are discussed, and suggestions to improve both classification design and application are offered.

5.1 Idealized Examples and Corpus Data

Raters were able to classify idealized examples with high accuracy. The use of accuracy scores allowed for the selection of competent raters. These raters were, however, not able to classify corpus data with a sufficient level of reliability. This indicates that classification results are not replicable between raters. As both the instructions, discussions of original text by the authors of the classifications, and rater selection ensured participants had ample understanding of and experience in classifying conditionals, the results of this study suggest that replication and generalization may be compromised by reliability issues.

To tackle this issue, it is suggested here that future classification design include not only clear, perhaps idealized examples of the types proposed, but also discussion of examples that seem to resist those very types. This will not only add reality value to the classification at hand, but instead of discouraging the actual use of the classification on real language data in following studies, it may also spawn further discussion of it, bringing our understanding of, in this case, conditionals as a whole forward, instead of licensing yet another, incompatible classification.

5.2 Types of Disagreement

Low reliability scores for the classification of conditionals may be the result of a number of problems (see also Sect. 5.2), but this does not mean that reliability should not be strived for. It is helpful to apply Spooren and Degand's (2010) distinction between two types of disagreement; first, disagreement can be a result of ambiguity, as language underspecifies meaning and context guides interpretation. Second, disagreement can be a result of a coding error, calling for an improved coding scheme. Ambiguity as a result of linguistic underspecification puts 'perfect agreement' out of reach (Spooren and Degand 2010, p. 251).

This study suggests that, while the interpretative nature of connections between antecedents and consequents of conditionals may indeed lead to lower agreement scores, the classifications discussed could be improved with respect to their representativeness of actual language use. A suggestion for future research therefore is to implement the principle of *total accountability* (McEnery and Hardie 2012, pp. 14–18), ensuring that classifications are not based on and illustrated by perfect examples, but by all data, including those examples that seem to resist clear classification.

5.3 Language Specificity

The low reliability scores reported in this study may be related to the fact that the items were taken from Dutch corpora, while the classification schemes are based mainly on English. Some types of conditional relations may be language specific. For example, in Dutch, the scalar type of Quirk et al.'s (1985) rhetorical conditional in (e.g., 'The package weighed ten pounds if it weighted an ounce.') is expressed more frequently by means of other constructions than a conditional.

To improve classification design, this study suggests further testing of existing classifications on languages beyond English, which may, in turn, enhance our understanding of cross-linguistic similarities and differences in the use of conditionals. In a similar vein, text dimensions such as mode (spoken, written) and register (formal, informal) should be incorporated into discourse analysis of conditionals, as conditionals may be used very differently in, for instance, argumentative discourse and instructional texts (cf. Reuneker 2020). This may lead to better task design for raters and, in turn, higher reliability and reproducibility of annotation results.

5.4 Linguistic Underspecification

Conditionals are linguistically underspecified for the types of relation between antecedent and consequent. Based on a corpus study of Dutch conditionals, Reuneker (2022a) considers the connections between antecedents and consequents of conditionals to be conversational implicatures (cf. Grice 1989), with high context-dependence, and only weak dependence on linguistic features. A relevant question is why raters seem not to be able to classify conditionals reliably, while in ordinary language use, conditionals seem to be interpreted largely without problems. One possible explanation is that the strict boundaries between types to be employed in classification tasks is not necessary in ordinary interpretation for communication to be successful.

One step that may have enriched the present study is a group discussion after the classification task. In post-annotation discussions, it may be expected that raters will reach consensus on specific examples, as they may provide arguments for their interpretation taking into account both context and world-knowledge. While this may impact the independence of the raters (see Sect. 3), a precondition for reliability, it may, in the end, produce better results for a specific study. A related suggestion for future research is to implement a 'think aloud' protocol to gain insight into the interpretation of raters (see e.g., Krueger and Casey 2009) and to enhance rater training, without compromising the independence of raters.

5.5 A Final Note

As conditionals enable us to express our thoughts about possible states of the world, they are indispensable cognitive aspects of our capabilities for reasoning and argumentation. In line with the hypothesis by Mercier and Sperber (2019), the use of

conditionals may be primarily argumentative. This reflects the findings by Evans (2005), who shows that conditionals are often interpreted as inducements or advice, and are understood primarily by their perlocutionary effect. Moreover, from the perspective of theories of argumentation, conditionals often have the status of a connecting premise (see e.g., van Eemeren and Snoeck Henkemans 2017, pp. 50–51), and can expressed to facilitate specific strategic manoeuvres.

While the role of conditionals in argumentation cannot be understated, their analysis in language use is impacted by issues of reliability. This study introduced novel ways of analyzing agreement, which may facilitate the identification of factors responsible for low reliability scores in the classification of conditionals and other, more interpretative features of argumentative language use.

Acknowledgements This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 023.005.085. Parts of this study have been previously published in Reuneker (2022a). I would like to thank the reviewers for their feedback, the participants of the European Conference on Argumentation (ECA) of 2022 in Rome for their suggestions, and the international jury of six leading argumentation scholars for awarding this article with the ECA Frans van Eemeren Prize 2022.

References

Artstein, R., and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 (4): 555–596.

Athanasiadou, A., and R. Dirven. 1997a. Conditionality, hypotheticality, counterfactuality. In *On conditionals again*, ed. A. Athanasiadou and R. Dirven, 61–96. Philadelphia: Amsterdam.

Athanasiadou, A., and R. Dirven. 1997b. Pragmatic conditionals. In Foolen, A., and Van Der Leek, F. (Eds.), *Constructions in Cognitive Linguistics: Selected papers from the Fifth International Cognitive Linguistics Conference, Amsterdam, 1997b* (pp. 1–26). Amsterdam: John Benjamins.

Austin, J. L. 1961. Ifs and cans. In Urmson, J. O., and Warnock, G. J. (Eds.), *Philosophical Papers*. Oxford: Oxford University Press.

Banerjee, M., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics* 27 (1): 3–23.

Bennett, J. 2003. *A philosophical guide to conditionals*. Oxford: Oxford University Press.

Bolognesi, M., R. Pilgram, and R. van den Heerik, 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 1984–2001.

Carter-Thomas, S., and E. Rowley-Jolivet. 2008. If-conditionals in medical discourse: From theory to disciplinary practice. *Journal of English for Academic Purposes* 7 (3): 191–205.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37–46.

Dancygier, B. 1998. *Conditionals and prediction: time, knowledge, and causation in conditional constructions*. Cambridge; New York: Cambridge University Press.

Dancygier, B., and E. Sweetser. 2005. *Mental spaces in grammar: Conditional constructions*. Cambridge: Cambridge University Press.

Declerck, R., and S. Reed. 2001. *Conditionals: a comprehensive empirical analysis*. Berlin; New York: Mouton de Gruyter.

Deygers, K., V. Van Den Heede, S. Grondelaers, D. Speelman, and H. Van Aken. 2000. Het CONDIV-corpus Geschreven Nederlands. *Nederlandse Taalkunde* 5 (4): 356–363.

Edgington, D. 2022. Suppose and tell: The semantics and heuristics of conditionals. *History and Philosophy of Logic* 43 (2): 188–195.

Evans, J.S.B.T. 2005. The social and communicative function of conditional statements. *Mind & Society* 4 (1): 97–113.

Evans, J.S.B.T., and D.E. Over. 2004. *If: Supposition, pragmatics, and dual processes*. Oxford: Oxford University Press.

Ferguson, G. 2001. If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes* 20 (1): 61–82.

Gabrielatos, C. 2010. *A corpus-based examination of English if-conditionals through the lens of modality: Nature and types*. PhD dissertation. Lancaster University.

Geis, M.L., and A.M. Zwicky. 1971. On invited inferences. *Linguistic Inquiry* 2 (4): 561–566.

Grice, H.P. 1989. *Studies in the way of words*. Cambridge, Massachusetts: Harvard University Press.

Gwet, K. L. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics, LLC.

Hartmann, S., and U. Hahn. 2020. 'A New Approach to Testimonial Conditionals'. In Denison, S., Mack, M., Xu, Y., and Amstrong, B. C., *Proceedings of the 42nd annual conference of the cognitive science society*, 981–986.

Hayes, A.F., and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1 (1): 77–89.

Kaplan, A., and J. M. Goldsen. 1965. The reliability of content analysis categories. *Language of I Politics*, 83–112.

Krueger, R., and M. Casey. 2014. *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage Publications.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

McEnery, T., and A. Hardie. 2012. *Corpus linguistics: method, theory and practice*. Cambridge, New York: Cambridge University Press.

Mercier, H., and D. Sperber. 2019. *The Enigma of reason*. Harvard: Harvard University Press.

Oaksford, M., and N. Chater. 2010. *Cognition and conditionals: Probability and logic in human thinking*. USA: Oxford University Press.

Oostdijk, N. 2000. Het corpus Gesproken Nederlands. *Nederlandse Taalkunde* 5 (3): 280–284.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. L. Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Retrieved from https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf on January 12, 2017.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive grammar of the english language*. London: Longman.

Reuneker, A. 2020. Clause order and syntactic integration patterns in Dutch conditionals. In *Linguistics in the Netherlands 37*, ed. E. Tribushinina and M. Dingemanse, 119–134. Amsterdam: John Benjamins Publishing.

Reuneker, A. 2022a. *Connecting Conditionals: A Corpus-Based Approach to Conditional Constructions in Dutch*. PhD thesis. Amsterdam: LOT/Netherlands Graduate School of Linguistics.

Reuneker, A. 2022b. *Data and scripts for 'Connecting Conditionals'.* DataverseNL. https://doi.org/10.34894/3QTEKH.

Rosch, E. 1978. Principles of categorization. In Lloyd, B. B., Wiles, J., and Rosch, E. (Eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Sanders, T.J., W.P. Spooren, and L.G. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15 (1): 1–35.

Sandri, G. 1969. On the logic of classification. *Quality & Quantity* 3 (1): 80–124.

Scholman, M. C. J., V. Demberg, and T. J. M. Sanders. 2022. Descriptively Adequate and Cognitively Plausible? Validating Distinctions between Types of Coherence Relations. *Discours 30*. https://doi.org/10.4000/discours.12075.

Spooren, W., and L. Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6 (2): 241–266.

Sweetser, E.E. 1990. *From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change*. Cambridge: Cambridge University Press.

Taylor, J.R. 2003. *Linguistic categorization*. Oxford: Oxford University Press.

Van der Auwera, J. 1986. Conditionals and speech acts. In Traugott, E. C., ter Meulen, A., Snitzer Reilly, J., and Ferguson, C. A. (Eds.), *On conditionals* (pp. 197–214). Cambridge: Cambridge University Press.

van Eemeren, F.H., and A.F. Snoeck Henkemans. 2017. *Argumentation: Analysis and evaluation*, 2nd ed. New York: Routledge.

van Gorp, B. 2007. The constructionist approach to framing: Bringing culture back in. *Journal of Communication* 57 (1): 60–78.

Verhagen, V., and M. Mos. 2016. Stability of familiarity judgments: Individual variation and the invariant bigger picture. *Cognitive Linguistics* 27 (3): 307–344.

Wierzbicka, A. 1997. Conditionals and counterfactuals: Conceptual primitives and linguistic universals. In *On conditionals again*, ed. A. Athanasiadou and R. Dirven, 15–60. Amsterdam: John Benjamins Publishing Company.