Surface and Contextual Linguistic Cues in Dialog Act Classification: A Cognitive Science View

Linders, Guido M ; Louwerse, Max M

DOI: https://doi.org/10.1111/cogs.13367

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13367

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13367

Surface and Contextual Linguistic Cues in Dialog Act Classification: A Cognitive Science View

Guido M. Linders,[a,b] Max M. Louwerse[a]

[a]*Department of Cognitive Science & Artificial Intelligence, Tilburg University*
[b]*Department of Comparative Language Science, University of Zurich*

Received 26 June 2022; received in revised form 26 September 2023; accepted 9 October 2023

Abstract

What role do linguistic cues on a surface and contextual level have in identifying the intention behind an utterance? Drawing on the wealth of studies and corpora from the computational task of dialog act classification, we studied this question from a cognitive science perspective. We first reviewed the role of linguistic cues in dialog act classification studies that evaluated model performance on three of the most commonly used English dialog act corpora. Findings show that frequency-based, machine learning, and deep learning methods all yield similar performance. Classification accuracies, moreover, generally do not explain which specific cues yield high performance. Using a cognitive science approach, in two analyses, we systematically investigated the role of cues in the surface structure of the utterance and cues of the surrounding context individually and combined. By comparing the explained variance, rather than the prediction accuracy of these cues in a logistic regression model, we found that (1) while surface and contextual linguistic cues can complement each other, surface linguistic cues form the backbone in human dialog act identification, (2) with word frequency statistics being particularly important for the dialog act, and (3) the similar trends across corpora, despite differences in the type of dialog, corpus setup, and dialog act tagset. The importance of surface linguistic cues in dialog act classification sheds light on how both computers and humans take advantage of these cues in speech act recognition.

Keywords: Dialog act classification; Speech act classification; Speech acts; Linguistic cues; Pragmatics

Correspondence should be sent to Guido M. Linders, Department of Comparative Language Science, University of Zurich, Affolternstrasse 56, 8050 Zurich, Switzerland. E-mail: guido.linders@uzh.ch

1. Introduction

The intention behind what we say does not always seem to be conveyed in the words we express. "Can you pass me the salt?" is not a question regarding one's salt-passing ability, and neither is the cashier's "how are you today?" a question to learn more about your inner feelings. Instead, a request is made here to pass the salt and the cashier's question is not much more than a greeting. Language users generally have no difficulty identifying the intentions behind an utterance. But how are they able to do this if the cues do not lie in the lexical information? What linguistic information do language users use to extract the intention behind an utterance?

An answer to the question what the sources of information are to identify the intentions behind an utterance starts with a classification of utterances and intentions. Austin (1962) proposed a distinction between locutionary, illocutionary, and perlocutionary acts, taking the first steps toward formalizing speaker intentions. The literal meaning of an utterance is the locutionary act, the intention behind the utterance is the illocutionary act, and the effect or outcome of the locutionary and illocutionary act is the perlocutionary act. Following Austin's distinction between locutionary, illocutionary, and perlocutionary acts, Searle (1976) proposed a categorization of the illocutionary acts, more commonly known as speech acts. Searle distinguished five mutually exclusive categories: assertives (e.g., "Mary kissed John"), directives (e.g., "Will Mary kiss John?"), commissives (e.g., "I will take you out for dinner tonight"), expressives (e.g., "I am thrilled I am going to get married!"), and declarations (e.g., "I, Mary, hereby take you, John, to be my husband").

These speech act categories are relevant theoretically, but the important question here is how a speech act can be identified by a language user, that is, how illocutionary forces map onto the locutionary utterances. An early proposal is that specific verbs mark the speech acts. Verbs such as "promise," "declare," "ask," "suggest," and "apologize" mark the intention of the speaker. Wierzbicka (1987) provided a detailed semantic breakdown of over a 100 English speech act verbs. Searle and Vanderveken (1985; but see also Levinson, 1983) proposed so-called illocutionary force indicating devices (IFIDs) that can help with the identification of a speech act. These IFIDs include linguistic markers in the surface structure of the utterance such as the word order, the mood of the verb, performative verbs, as well as extralinguistic cues such as stress and intonation contours. However, the problem with speech acts is that utterances may not always be marked with IFIDs and that there is no fixed mapping between speech acts and utterances. The utterance "I will come to your house" could be an assertive (a general statement that I will come to your house), a directive (if the sentence is followed by a question mark or a rising pitch), but it could equally well be a threat, a promise, or an expressive. And conversely, the same speech act can be expressed in different utterances ("I will take you out for dinner," "I promise I will take you out for dinner," "I will most definitely take you out for dinner"). Such utterances are thus underspecified. In other words, surface linguistic cues such as IFIDs may not help to identify speech acts.

Whereas Austin (1962), Searle (1976), and Wierzbicka (1987) focused on the speech act of a single utterance, Schegloff and Sachs (1973) emphasized the interactional function of speech acts by considering utterance pairs as the unit of communication. These adjacency

pairs included both the utterance of a speaker and the response by the other speaker, and provided some potentially critical contextual cues. "How are you today?" could be an inquiry of your inner feelings, but when followed by a "fine, thanks" it is not. But just like the problem of identifying a speech act based on the surface linguistic information, it is not clear to what extent the linguistic context adds to the identification of a speech act.

In line with Schegloff and Sachs (1973), Clark (1996) also emphasized that speech acts are joint acts, but, moreover, argued that they are tightly interlinked with the linguistic cues underlying the speech act. Dialog participants need to have common agreement on the meaning of communicative signals. A speech act can only be understood if the communicative signals are conventionalized by all participants involved. In addition, many signals will only get a conventional meaning attached when put into context, thereby relying on the dialog history, the common ground, and the situation. Clark (1996) devised an action ladder, which consists of four different steps to describe a joint action at different levels. At the lowest level, a speaker executes a (verbal or nonverbal) behavior to which the listener pays attention. One level up the ladder, the executed behavior is presented as a signal that the listener tries to identify. Moving another level up, the signal is described in terms of its (semantic) meaning that the listener tries to interpret. Finally, at the highest level, the signal is described as the proposal or communicative intention which the listener tries to consider. In other words, this action ladder is a framework on how linguistic signals are related to the communicative intention. It can, furthermore, be seen as an extension of Austin's (1962) distinction between locutionary, illocutionary, and perlocutionary acts. However, where Austin focused just on describing the speaker's action, Clark included the perceiver as well, building an action ladder of joint acts.

Subsequent research has primarily deviated from defining speech acts, and focused on specific processes involved in speech act recognition, such as conversational implicatures (Grice, 1975; Levinson, 2000), presuppositions (Beaver, 1997; Stalnaker, 2002), and figurative language, including metaphors and irony (Haverkate, 1990; Mac Cormac, 1985; Wilson, 2006). Pragmatic inferences of, for example, conversational implicatures, metaphors, and irony have been formalized in a Bayesian computational framework: rational speech acts (Goodman & Frank, 2016; Kao & Goodman, 2015; Kao, Bergen, & Goodman, 2014). The goal of this framework is to better understand the psychological mechanisms behind different pragmatic inferences (Goodman & Frank, 2016). Based on the seminal work by Grice (1975), who argued that conversational participants behave according to a set of rationales (i.e., maxims of conversation), one of the central hypotheses of the framework was that speakers and listeners recursively reason about the mental state of each other, assuming a fully rationally behaving agent, and using probabilistic Bayesian inferences to infer the intended meaning of an utterance.

Studies on how linguistic cues signal speech acts are overall theoretical in nature. Some studies have, however, investigated the neural and cognitive processes involved in speech act recognition, but they typically look at very broad distinctions of speech acts, such as direct versus indirect speech acts (Boux, Margiotoudi, Dreyer, Tomasello, & Pulvermüller, 2023; Licea-Haquet, Reyes-Aguilar, Alcauter, & Giordano, 2021; Tromp, Hagoort, & Meyer, 2016; see Tomasello, 2023 for an overview). Importantly, the recognition of speech acts seems to happen automatic (Holtgraves, 2008) and seems to happen early on in the utterance to

facilitate fast turn-taking in dialog (Gisladottir, Chwilla, & Levinson, 2015). Language users anticipate for the speech act, likely already before the utterance is uttered (Boux, Tomasello, Grisoni, & Pulvermüller, 2021; Gisladottir, Bögels, & Levinson, 2018). Differences in behavioral and neural indicators have been identified for indirect and direct speech acts, for example, in the cognitive processes involved with different speech acts processed differently in the brain (Boux et al., 2023).

Some studies have, furthermore, empirically investigated the communicative signals that are involved in the identification of speech acts. These include acoustic signals (Hellbernd & Sammler, 2016; Ruytenbeek, Bergen, & Trott, 2023; Trott, Reed, Kaliblotzky, Ferreira, & Bergen, 2023), nonverbal signals, such as gestures (Bucciarelli, Colle, & Bara, 2003; Enrici, Adenzato, Cappa, Bara, & Tettamanti, 2011), facial expressions (Domaneschi, Passarelli, & Chiorri, 2017) and eye movements (Jang, Mallipeddi, Lee, Kwak, & Lee, 2014), and context (Abbeduto, Furman, & Davies, 1989; Gibbs, 1983). These studies have primarily focused on extralinguistic cues, such as nonverbal and acoustic cues. Specific cues in the linguistic contribution or context have hardly ever been investigated in speech act recognition.

The fact that surface linguistic cues have not played a prominent role in research on speech acts is perhaps not surprising. Surface linguistic cues are hard to formalize and empirically validate and, as we have seen, do not constitute the only cue type that is important for the speech act (illocutionary act). Yet, it is unclear how these linguistic cues support the identification of speech acts. That question is not only relevant for cognitive and social psychology, psycholinguistics, and neuroscience, but also for computational linguistics.

Take, for instance, one application in computational linguistics, that of embodied conversational agents (Graesser et al., 2004). When these agents need to respond to the incoming utterance, it is critical that they do not only identify the meaning of the utterance, but also the speech act behind the utterance to respond in a natural way. For example, in intelligent tutoring systems, a student stating "I don't know" or "I am lost" is requesting information from the system, and capturing the intention behind these utterances is critical. In the computational linguistics literature, these kinds of speech acts are generally referred to as *dialog acts*, highlighting their use in conversational interactions. Dialog acts mark the function of an utterance in the dialog (Bunt, 1994) and are, therefore, comparable to the more general speech acts that are not reserved for just dialog.[1]

In computational linguistics, human speech act recognition is framed as a computational problem and referred to as dialog act classification. Computational algorithms are trained on dialog act corpora containing transcriptions of natural interactions between several dialog participants, where each utterance is annotated with a dialog act by independent annotators. Using cues extracted from the utterances and the surrounding context, the algorithm then learns which cues generally indicate which dialog act, and the trained model is then used to infer the dialog act of an unseen utterance. The goal is to find the combination of features (i.e., computational implementations of cues), together with the learning algorithm that leads to the highest performance. Since these algorithms are trained on naturalistic interactions, they provide valuable insights into the cues that might be important in human speech act recognition.

The results from the field of dialog act classification can be especially interesting for studying the underlying mechanisms of human dialog act identification for multiple reasons. First, the availability of large corpora with naturalistic conversations, annotated with speaker intentions, allows for thorough investigations of the (cognitive) mechanisms involved. Furthermore, the wealth of studies in dialog act classification have used a large array of cues and approaches for predicting dialog acts. While these approaches and cues are essentially computational in nature, they are based on statistical patterns in naturalistic conversations and can, therefore, be a proxy for understanding the mechanisms underlying the identification of speech acts.

While dialog act corpora are a valuable source of information for the investigation of speech acts, they also have several limitations. Annotations are oftentimes based on the transcriptions only (Louwerse, et al., 2012), with no acoustic cues available to the annotator (Jurafsky, Shriberg, & Biasca, 1997). Moreover, given the human ability to understand the intention of the speaker and assuming that surface and contextual linguistic cues would aid dialog act classification, one would perhaps expect near-perfect inter-rater reliability scores. However, the contrary is true. For example, the inter-annotator agreement, quantified using Cohen's Kappa, was 0.80 on the Switchboard corpus (Jurafsky et al., 1997), 0.80 on the ICSI Meeting Recorder Dialog Act (MRDA) corpus (Shriberg, Dhillon, Bhagat, Ang, & Carvey, 2004), and 0.83 on the HCRC Map Task corpus (Carletta et al., 1997), which have tagsets consisting of 42, 6, and 13 dialog acts, respectively. Finally, typically a very small portion of the data is used for evaluating the performance of the models. These test sets are often standardized and used throughout the different studies, potentially obfuscating the resulting findings.

The aim of the current paper is to better understand the role of the surface and contextual linguistic cues underlying the identification of human speech acts, using the results and corpora from the computational task of dialog act classification. The goal here is to take a cognitive science approach, understanding the mechanisms of speech act recognition *using explanation*, rather than a computational approach, *maximizing the prediction* accuracy (Jones, 2017; McClelland, 2009).

There are multiple reasons why the study of the mechanisms behind human speech act recognition can benefit from the results of dialog act classification. First, as already mentioned, the availability of large dialog act corpora can provide important information on the statistical tendencies between the linguistic cues and speech acts, although limited to symbolic information only. In empirical studies involving humans, acquiring large amounts of language data is both very costly and time-consuming. Second, the different dialog act classification studies provide important indications on which cues are important for human speech act recognition, something which is not well-investigated in the cognitive science literature. And finally, while it is hard to study the role that IFIDs and linguistic cues play through empirical research, computational models can seamlessly utilize such information.

Vice versa, there are also several reasons why it is valuable to investigate dialog act classification using a cognitive science approach. First, dialog act classification models have increasingly relied on more complex learning algorithms, making it increasingly hard to understand the role of different cues in human speech act recognition (Jones, 2017; McClelland, 2009). Especially in cognitive science, a simpler explanation and a less complex model is usually

preferred (Chater & Vitányi, 2003). Consequently, the current study follows the principle of parsimony to gain insight into the mechanisms (explanation) over accuracy (prediction). Second, while the focus on prediction is valuable, combining results from explanation and prediction is important in building theories on human behavior (Rosenberg, Casey, & Holmes, 2018; Shmueli, 2010; Tinga, Kuperus, Carvalho, & Louwerse, 2019). Explanation is often poorly distinguished from prediction (Shmueli, 2010; Yarkoni & Westfall, 2017), but offers additional valuable information on the causal relationships between independent variables (cues) and the dependent variable (observed dialog act), which prediction cannot provide as the statistical model is always part of the equation. Explanatory models are, furthermore, the default in cognitive science. Changing from the computational paradigm to a cognitive one, we can treat the cues established in the dialog act classification literature as our independent variables, observed from empirical data on human−human communication (dialog act corpora). Hence, these dialog act corpora seem not only suited for computational studies on predicting unseen dialog acts, but also for understanding the role of different cues in explaining the dialog act.

In part one of the current study, we review the role of linguistic information in dialog act classification systems. We investigate the cues that have been used in 50 different dialog act classification studies and summarize them into different categories with the goal of gaining insight into the linguistic features that contribute to their performance on three of the most commonly used English dialog act corpora. In the second part, we investigate the role of cues in the surface structure of the utterance and cues of the surrounding context both individually and in different combinations ourselves.

2. Identifying cues in existing dialog act classification studies

We reviewed 50 representative dialog act classification models in 44 publications that reported an accuracy on one of three English dialog corpora, the Switchboard Dialog Act corpus, the ICSI MRDA corpus, and the Map Task corpus. These corpora are not only the ones most commonly used in dialog act classification, but are also used for further analyses in the second part of this paper. In the 50 dialog act classification studies, we identified (1) the approach they used, (2) the cues they used to reach the performance, (3) the corpus (or corpora) used in the evaluation, and (4) the performance of these models. An overview of the 50 dialog act classification studies is given in Table 1. The studies have been ordered on their approach taken and (subsequently) on the year they were published.

Before we discuss these studies, it is important to briefly mention the different cue types and types of studies that were included in this overview. The linguistic cues that the dialog act classification studies use can generally be classified into surface, contextual, and acoustic cues. Surface cues provide cues through lexical, syntactic, or (distributional) semantic information such as "do you" at the start of an utterance indicating that the speaker is requesting new information. Contextual cues provide cues through information from the context around the utterance, for example, through previous dialog acts or surface information in previous utterances that provide hints on what the current dialog act could be. For instance, if the

Table 1
Overview of 50 dialog act classification studies ordered by approach

No.	Study	Approach				Surface linguistic cues									Contextual linguistic cues						Accuracy		
		Frequency-based	Machine learning	Deep learning	Sequence-based	Character sequences	Pretrained character embeddings	Word sequences	PoS sequences	Utterance length	Other syntactic cues	Noncontextualized word embeddings	Contextualized word embeddings	Latent semantic analysis	True dialog act sequences	Predicted dialog act sequences	Surface cues of context	Speaker cues	Utterance position in turn	Discourse organization cues	Switchboard corpus	MRDA corpus	Map Task corpus
1	Garner, Browning, Moore, and Russell (1996)	•						•								•							54.8
2	Stolcke et al. (2000)	•			•			•								•					71.0		
3	Ji and Bilmes (2005)	•			•			•								•						80.3	
4	Webb, Hepple, and Wilks (2005)	•						•		•											69.1		
5	Louwerse and Crossley (2006)	•						•										•					58.1
6	Lager and Zinovjeva (1999)		•					•			•				•			•					62.1
7	Grau, Sanchis, Castro, and Vilar (2004)		•					•													66.0		
8	Ang, Liu, and Shriberg (2005)		•					•		•												79.5	
9	Ji and Bilmes (2006)		•		•			•								•						81.5	
10	Surendran and Levow (2006)		•		•			•								•							59.1
11	Verbree, Rienks, and Heylen (2006)		•					•	•	•	•				•						65.7	89.3	
12	Novielli and Strapparava (2009)		•						•					•							68		
13	Sridhar, Bangalore, and Narayanan (2009) (1)		•					•	•		•				•			•			76.0		75.2
14	Sridhar et al. (2009) (2)		•					•	•		•						•	•			72.0		69.9
15	Di Eugenio, Xie, and Serafin (2010)		•						•					•				•		•			78.8
16	Ribeiro et al. (2015) (1)		•					•							•						79.1		
17	Ribeiro et al. (2015) (2)		•					•									•				74.9		
18	Brychcín and Král (2017)		•		•			•				•				•					72.9		
19	Kalchbrenner and Blunsom (2013)			•				•									•				73.9		
20	Ji, Haffari, and Eisenstein (2016)			•				•								•	•				77.0		
21	Khanpour, Guntakandla, and Nielsen (2016)[2]			•								•									75.8	86.8	
22	Lee and Dernoncourt (2016)			•								•					•				73.1	84.6	
23	Li and Wu (2016)			•	•					•		•					•	•	•		79.4		
24	Shen and Lee (2016)			•				•									•				72.6		
25	Liu, Han, Tan, and Lei (2017)			•	•							•					•	•			77.2		

(*Continued*)

Table 1
(Continued)

No.	Study	Approach				Surface linguistic cues									Contextual linguistic cues						Accuracy		
		Frequency-based	Machine learning	Deep learning	Sequence-based	Character sequences	Pretrained character embeddings	Word sequences	PoS sequences	Utterance length	Other syntactic cues	Noncontextualized word embeddings	Contextualized word embeddings	Latent semantic analysis	True dialog act sequences	Predicted dialog act sequences	Surface cues of context	Speaker cues	Utterance position in turn	Discourse organization cues	Switchboard corpus	MRDA corpus	Map Task corpus
26	Ortega and Vu (2017)			•								•					•				73.8	84.3	
27	Papalampidi, Iosif, and Potamianos (2017)			•				•				•					•				75.6		
28	Tran, Zukerman, and Haffari (2017a)			•	•			•								•	•				74.2		65.9
29	Tran, Zukerman, and Haffari (2017b)			•	•			•								•					75.6		62.9
30	Bothe, Magg, Weber, and Wermter (2018) (1)			•			•					•									71.2		
31	Bothe, Magg et al. (2018) (2)			•			•					•					•				77.4		
32	Bothe, Weber et al. (2018)			•			•										•	•			77.3		
33	Cerisara, Král, and Lenc (2018) (1)			•				•									•				72.8		
34	Cerisara et al. (2018) (2)			•								•					•				72.5		
35	Chen, Yang, Zhao, Cai, and He (2018)			•	•	•			•			•					•				81.3	91.7	
36	Duran and Battle (2018)			•				•													75.5		
37	Kumar, Agarwal, Dasgupta, and Joshi (2018)			•	•							•					•				79.9	90.2	
38	Ravi and Kozareva (2018)[3]			•	•	?		?													83.1	86.7	
39	Wan et al. (2018)			•								•					•				81.5		68.5
40	Li, Lin, Collinson, Li, and Chen (2019)			•	•		•					•					•				82.3	92.2	
41	Raheja and Tetreault (2019)			•	•		•					•	•				•				82.9	91.1	
42	Ribeiro, Ribeiro, and de Matos (2019a)			•		•						•									79.0		
43	Ribeiro, Ribeiro, and de Matos (2019b)			•		•							•				•	•			79.1	90.6	
44	Zhao and Kawahara (2019) (1)			•	•			•									•				79.0		
45	Zhao and Kawahara (2019) (2)			•	•			•									•	•			80.2		
46	Colombo et al. (2020)			•	•							•					•				85.0	91.6	
47	Dai et al. (2020) (1)			•				•									•				78.9		
48	Dai et al. (2020) (2)			•	•			•									•				80.3		
49	Duran, Battle, and Smith (2021)			•									•								76.2		62.9
50	Yano, Tamura, Ninomiya, and Obayashi (2021)			•	•			•									•		•		78.1		

Note. The numbers in the first columns are merely used as a reference later in this paper.

previous utterance was a question by another speaker, it provides a cue that the speaker likely will address this question in the current utterance. Acoustic cues provide hints on pitch and stress patterns in an utterance, such as a pitch rise at the end of an utterance indicating a question is being asked. Surface and contextual cues can be extracted from the transcriptions of the dialog, whereas acoustic cues can only be derived from the actual speech signals.

Given the wealth of different surface and contextual cues and following the lion share of dialog act classification studies, we only focus on linguistic cues that can be extracted from the transcriptions. We hence disregard the acoustic cues in the current study. Even though acoustic cues have shown to be a valuable source of information for dialog act classification systems (Hoque, Sorower, Yeasin, & Louwerse, 2007; Sridhar et al., 2009; Stolcke et al., 2000) as well as in human speech act recognition (Hellbernd & Sammler, 2016; Ruytenbeek et al., 2023; Trott et al., 2023), the number of computational studies that use acoustic cues is relatively small with most dialog act classification systems focusing on transcribed dialog. In addition, by focusing on transcribed speech, these studies avoid the problem of segmenting speech into turns and utterances.

2.1. Approaches to dialog act classification

In addition to the distinction between surface and contextual cues being used in the dialog act classification studies, three different classification approaches can be distinguished: (1) frequency-based approaches, (2) machine learning approaches, and (3) deep learning approaches. Early dialog act classification studies often used a frequency-based approach to classify dialog acts. Such models were mainly based on count statistics of words using *n*-grams, often in combination with a contextual model based on dialog act count statistics of the previous dialog acts. Importantly, frequency-based approaches, unlike the other approaches, do not include a learning algorithm. All five dialog act classification studies using a frequency-based approach that we reviewed (Table 1, Studies 1–5) made use of word sequence statistics in the form of word *n*-grams to represent utterances. While one of the reviewed studies did not make use of contextual cues (Webb et al., 2005), another study used information on who is the speaker as contextual cues (Louwerse & Crossley, 2006), and yet other studies represented context through incorporating information on the dialog acts of the previous utterances (Garner et al., 1996; Ji & Bilmes, 2005; Stolcke et al., 2000).

The advantage of frequency-based models is their simplicity, which is also their downside. Simplicity allows for an easy investigation of the statistical considerations of dialog acts, but prevents easy integration of different types of cues. This simplicity and the lack of the ability to optimize the model through a learning procedure, however, led to the abandonment of these models, when more sophisticated machine learning models became mainstream. Because of their flexibility, the range of models that use a machine learning approach is more varied (Table 1, Studies 6–18). Machine learning models have made use of many different learning algorithms, including naive Bayes (Grau et al., 2004), logistic regression (Ang et al., 2005; Sridhar et al., 2009), support vector machines (Ribeiro et al., 2015; Surendran & Levow, 2006), decision trees (Verbree et al., 2006), Bayesian networks (Brychcín & Král, 2017; Ji

& Bilmes, 2006), latent semantic analysis (Di Eugenio et al., 2010; Novielli & Strapparava, 2009), and transformation-based learning (Lager & Zinovjeva, 1999).

Because of their flexibility, machine learning approaches allow for multiple different cues. While word sequences, encoded through a bag-of-words approach, are dominant (e.g., Grau et al., 2004; Ribeiro et al., 2015; Sridhar et al., 2009; Surendran & Levow, 2006), syntactic cues (Di Eugenio et al., 2010; Lager & Zinovjeva, 1999; Novielli & Strapparava, 2009; Verbree et al., 2006) and semantic cues, through a latent semantic analysis (Di Eugenio et al., 2010; Novielli & Strapparava, 2009), have also been used. Some machine learning studies did not consider context (Ang et al., 2005; Grau et al., 2004; Novielli & Strapparava, 2009), but most others used contextual cues, such as surface information of previous utterances (Ribeiro et al., 2015; Sridhar et al., 2009), cues on the speakers of the utterances (Di Eugenio et al., 2010; Lager & Zinovjeva, 1999; Sridhar et al., 2009), and cues related to the organization of the discourse, through encoding turns with a hierarchical structure, such as subdialogs (Di Eugenio et al., 2010).

Most of the machine learning studies integrated surface and contextual cues simultaneously in a single feature vector. However, some used a hierarchical structure through Hidden Markov Models, where surface and contextual information are not presented to the model simultaneously (Brychcín & Král 2017; Ji & Bilmes, 2006; Surendran & Levow, 2006). Generally, models including some form of contextual cues performed better than their counterparts that only used an utterance encoding. But the increase in performance is, however, rather small (see Ribeiro et al., 2015; Sridhar et al., 2009 for a comparison, also presented in Table 1).

In machine learning models, there is a lot of flexibility with regard to learning procedures and encoding of cues, and the models can be simple enough to allow for investigating which cues are most predictive. However, at the same time, the model structures and encoding of cues in machine learning approaches are still rather simple, making it hard to move beyond the linear integration of cues and adopt more complicated hierarchical structures. Deep learning approaches address most of these issues (Table 1, Studies 18–50). Deep learning approaches to dialog act classification gained traction in recent years and have enhanced performance with models approaching or even surpassing the human inter-annotator agreement (Colombo et al., 2020; Raheja & Tetreault, 2019; Ravi & Kozareva, 2018). These models have a more hierarchical structure and typically encode the utterance and context at different levels. The utterance is often directly encoded in a recurrent neural network (e.g., Dai et al., 2020; Ji et al., 2016; Khanpour et al., 2016; Shen & Lee, 2016), or through first creating an embedding layer through word embeddings (e.g., Chen et al., 2018; Khanpour et al., 2016; Lee & Dernoncourt, 2016; Raheja & Tetreault, 2019), and more recently, also through character embeddings (Bothe, Weber et al., [2018]; Ribeiro et al., 2019a). Multiple utterance representations are sometimes used simultaneously, such as character sequences and pretrained word embeddings (Ribeiro et al., 2019a, 2019b), both pretrained character and word embeddings (Bothe, Magg et al., 2018) or part-of-speech (PoS) sequences, character embeddings, a named entities encoding, and pretrained word embeddings (Chen et al., 2018). Context is either modeled through a condensed vector of contextual cues (e.g., Chen et al., 2018; Colombo et al., 2020; Kumar et al., 2018; Raheja & Tetreault, 2019) or directly in the neural network structure

Table 2
Average and standard deviation of the accuracies of studies in the literature overview summarized per approach and corpus

	Switchboard		MRDA		Map Task	
Approach	*M*	*SD*	*M*	*SD*	*M*	*SD*
Frequency-based	70.05	1.34	80.30	4.44	58.33	8.05
Machine learning	72.06	4.62	83.43	23.48	60.60	18.61
Deep learning	75.58	3.51	87.52	3.06	64.40	6.52

with a node for each utterance (e.g., Cerisara et al., 2018; Kalchbrenner & Blunsom, 2013; Ortega & Vu, 2017; Ribeiro et al., 2019b). Some studies that used deep learning for dialog act classification did not include contextual cues (e.g., Duran & Battle, 2018; Duran et al., 2023; Khanpour et al., 2016; Ribeiro et al., 2019a), but most encoded context through using a condensed surface encoding of the previous utterances (e.g., Chen et al., 2018; Kumar et al., 2018; Yano et al., 2021; Zhao & Kawahara, 2019). Most deep learning models have regarded dialog act classification as classifying a sequence of dialog acts, without paying attention to the speaker or the structure of utterances into turns. Some more recent deep learning approaches have seen success with incorporating some form of incorporating speaker cues (e.g., Bothe, Weber et al., [2018]; Li & Wu, 2016; Ribeiro et al., 2019b; Zhao & Kawahara, 2019) and incorporating the turn structure (Li & Wu, 2016; Yano et al., 2021). Finally, many of the more recent deep learning models included some form of an attention mechanism (e.g., Li et al., 2019; Raheja & Tetreault, 2019; Tran et al., 2017a; Wan et al., 2018).

Deep learning models constitute the state-of-the art and their complex and flexible structures allow for hierarchical fusion of different cues. The downside of these models is that this added complexity makes the models harder to interpret and understand, as one cannot easily determine how each cue impacts the classification accuracy. Consequently, it is also difficult to gain insight into how to improve the performance of deep learning approaches to dialog act classification.

We have summarized all accuracies reported in the selection of studies of the literature in Table 1 into averages and standard deviations, per corpus and approach, presented in Table 2. This table shows that the differences in accuracy between frequency-based, machine learning, and deep learning approaches are relatively small, whereas the differences between corpora are considerable. Apparently, it is relatively easy to reach a reasonable performance, but very hard to reach a very high performance, even with the best-performing models. Importantly, these differences in accuracy for each approach are likely determined by the setup. For example, taking a sequence-based approach, as explained next, allows for better optimization of the sequence, and in turn, a potentially higher performance, because these models have the whole conversation to their availability (Dai et al., 2020). Some studies also made use of information on the true, gold-standard dialog act, which similarly could influence the performance, although it has generally been used to present a hypothetical upper bound of the performance (Ribeiro et al., 2015; Sridhar et al., 2009). In addition, some studies fully or partially removed

punctuation marks (e.g., Louwerse & Crossley, 2006; Wan et al., 2018; Webb et al., 2005), while others kept these markings (e.g., Ji et al., 2016; Kalchbrenner & Blunsom, 2013; Li & Wu, 2016; Papalampidi et al., 2017), a choice that can have a significant impact on the performance (Duran et al., 2023; Ribeiro et al., 2019a).

Given that the differences across the three approaches are small, the question remains what explains the differences in performance across the approaches. Perhaps the answer to this question lies less in the approach taken — frequency-based, machine learning, or deep learning — and more in the linguistic cues that the different dialog classification algorithms use.

2.2. Instance-based versus sequence-based

When we look at Table 1, we can identify two general approaches when classifying utterances into dialog acts in studies that have demonstrated automated dialog classification performance: those that classify a single utterance into a single dialog act (e.g., Cerisara et al., 2018; Ribeiro et al., 2015; Webb et al., 2005), and those that consider multiple utterances at the same time to classify them into dialog acts (e.g., Li et al., 2019; Stolcke et al., 2000; Surendran & Levow, 2006). The instance-based or utterance-based classification method is generally used for online processing, such as in dialog systems (Ahmadvand et al., 2019; Král, Cerisara, & Klečkova, 2006). After all, in online settings, only a single utterance (and the dialog history, but not the future utterances) is all that is available from a user at each classification step. The sequence-based classification method, on the other hand, classifies a sequence of all dialog acts at once, usually, all dialog acts in an entire conversation. This approach is, therefore, less useful for interactive systems because the system can obviously not wait until the last dialog act of the conversation before starting to predict an earlier one. Sequence-based classification methods, however, are very useful for creating corpora with tagsets, which can be used for training and testing purposes. Because of the availability of more (future) information when classifying an utterance, sequence-based approaches have the potential to reach a higher performance, as demonstrated in Dai et al. (2020) and shown in Table 1.

As also shown in Table 1, sequence-based models are present in all three approaches and are generally used for more advanced integration of context. But both frequency-based and machine learning approaches tend to rely more on an instance-based than a sequence-based approach. Frequency-based and machine learning models that make use of a sequence-based approach make use of Hidden Markov Models and dynamic programming algorithms to optimize the sequence of predicted dialog acts (Ji & Bilmes, 2005, 2006; Stolcke et al., 2000; Surendran & Levow, 2006). Sequence-based deep learning approaches typically encode context by connecting the utterances in a neural network structure (e.g., Colombo et al., 2020; Chen et al., 2018; Kumar et al., 2018; Raheja & Tetreault, 2019). Instance-based deep learning approaches, on the other hand, generally incorporate a vector representation of the previous utterances (e.g., Cerisara et al., 2018; Kalchbrenner & Blunsom, 2013; Ortega & Vu, 2017; Ribeiro et al., 2019b).

The distinction between instance- and sequence-based approaches is important to consider, because sequence-based models provide a non-naturalistic setting from a cognitive

perspective. Humans cannot rely on information of the future to identify the current dialog act, nor identify a long sequence of dialog acts at once. For applications of dialog act classification, such as embodied conversational agents, the same is true. Because the sequential information may convolute the understanding of linguistic cues in dialog act classification, we disregarded sequence-based models beyond this literature overview.

2.3. Surface and contextual linguistic cues

The range of cues, which can be broadly divided into cues related to the surface structure of the utterance and cues related to the surrounding context, is varied. Surface cues at almost any linguistic level, ranging from character to semantic level, have been successfully used in dialog act classification. Similarly, many different contextual aspects, ranging from predicted speaker intentions of previous utterances to cues on who is or was the speaker of the current or previous utterances, have been successfully used. But while all studies make use of one or more surface cues, not all studies make use of contextual cues (e.g., Ang et al., 2005; Duran & Battle, 2018; Duran et al., 2023; Novielli & Strapparava, 2009).

To provide a structured overview of the surface linguistic cues and the contextual linguistic cues in the 50 studies, we categorized the findings in Table 1 further in Fig. 1. This categorization closely follows the classification by Král, Pavelka, and Cerisara (2008) with a similar categorization of the linguistic cues into surface and contextual cues. The surface cues were further categorized into different linguistic levels: character, word, syntactic, and semantic, while the contextual cues were further divided into different types: dialog acts, surface cues from previous utterances, speaker cues, cues on the utterance position in the turn, and cues related to the discourse organization.

Character sequences encode utterances on a character level, thereby preserving the regularities in word inflections. Pretrained character embeddings have also been used in several deep learning models (e.g., Bothe, Magg et al., 2018; Bothe, Weber et al., [2018]; Li et al., 2019; Raheja & Tetreault, 2019). Both character sequences and pretrained character embeddings have primarily been used in addition to a semantic level cue (e.g., Chen et al., 2018; Li et al., 2019, Ribeiro et al., 2019a, 2019b), with the argument of also preserving statistical regulations within words.

Word sequences are the most frequently used cue (e.g., Dai et al., 2020; Grau et al., 2004; Kalchbrenner & Blunsom, 2013; Stolcke et al., 2000), likely because many dialog act classification frameworks heavily rely on lexical information (Duran & Battle, 2018; Jurafsky et al., 1997; Louwerse & Crossley, 2006). In frequency-based and machine learning approaches, word sequences have typically been encoded using *n*-grams (e.g., Garner et al., 1996; Grau et al., 2004; Ribeiro et al., 2015; Louwerse & Crossley, 2006). Deep learning methods have typically used recurrent neural networks to encode such sequences (e.g., Dai et al., 2020; Ji et al., 2016; Tran et al., 2017b; Zhao & Kawahara, 2019). Longer sequences of words preserve syntactic information in the form of word order. It is important to consider the fact that in sequences, structural information cannot be discerned from the information type (i.e., character, word, and PoS tags) anymore.

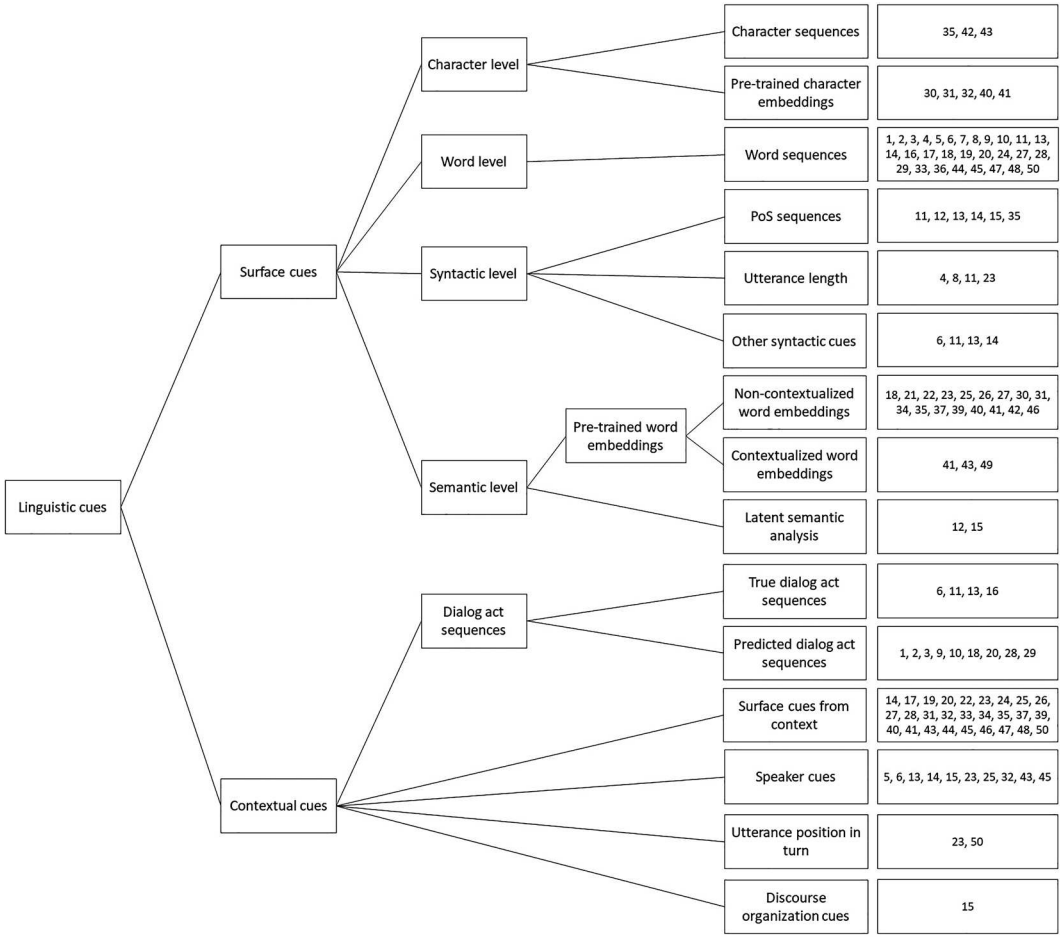

Fig. 1. Tree structure with a categorization of cues used in dialog act classification literature.
Note. To allow for identifying the studies that use these cues, we refer to their numbering in Table 1.

Syntactic cues solely capture the structure, and not the contents of an utterance, using, for example, PoS *n*-grams (Verbree et al., 2006) or the length of the utterance (Lager & Zinovjeva, 1999; Li & Wu, 2016; Verbree et al., 2006; Webb et al., 2005). Other syntactic cues have been used as well, such as labels representing elementary constituent trees (Sridhar et al., 2009) or labels representing the general syntactic structure of the whole utterance (Novielli & Strapparava, 2009). Syntactic information has primarily been used in machine learning approaches.

Finally, we identified a semantic level. Here, word information is condensed into a vector using different dimensionality reduction techniques. For example, some machine learning and many deep learning studies used pretrained noncontextualized word embeddings, such as Word2Vec or GloVe (e.g., Brychcín & Král, 2017; Colombo et al., 2020; Kumar et al.,

2018; Lee & Dernoncourt, 2016). Here, words are represented by a dense vector containing (distributional) semantic information. Noncontextualized distributional semantic approaches to encoding surface information might be suboptimal, because dialog acts seem to rely more on the statistical regulations of words in an utterance than on the semantic interpretation of the utterance (Cerisara et al., 2018; Duran & Battle 2018; Milajevs & Purver, 2014). More recently, also contextualized word embeddings, such as BERT (Duran et al., 2023; Ribeiro et al., 2019b) have been used in deep learning models. The main difference between noncontextualized and contextualized word embeddings is that contextualized embeddings encode full sentences at once, thus preserving word order information, and subsequently, the meaning of the whole sentence, while noncontextualized embeddings encode individual words, thereby fully relying on the semantic interpretation of individual words. Finally, some machine learning approaches have incorporated the latent semantic analysis dimensionality reduction technique to capture the semantic similarity between utterances (Di Eugenio et al., 2010; Novielli & Strapparava, 2009).

Especially in earlier studies, the dialog acts of previous utterances were used as contextual cues in the prediction of the current dialog acts. While some studies have used the true, that is, gold-standard, dialog acts of previous utterances in their model (Lager & Zinovjeva, 1999; Ribeiro et al., 2015; Sridhar et al., 2009; Verbree et al., 2006), other studies have used the dialog acts predicted by the model itself to serve as information for the prediction of the next utterance (e.g., Ji & Bilmes, 2006; Stolcke et al., 2000; Surendran & Levow, 2006; Tran et al., 2017a).

More recently, studies have used an utterance representation of the previous utterances which serves as information in the prediction. While machine learning approaches have encoded this type of information using word *n*-grams (Ribeiro et al., 2015) or a combination of word *n*-grams, PoS tags, and other syntactic features (Sridhar et al., 2009), this cue is more naturally used in deep learning approaches, where it has typically been directly encoded in the recurrent structure of the model (e.g., Cerisara et al., 2018; Dai et al., 2020, Kalchbrenner & Blunsom, 2013; Kumar et al., 2018).

Speaker information can be useful in dialog where both speakers have a different role, as, for example, is the case in the Map Task corpus (Louwerse & Crossley, 2006). These cues can, moreover, be useful when other contextual cues are used in the prediction (Zhao & Kawahara, 2019). Since turns can consist of multiple utterances, it is not always directly obvious which cues belong to which speakers. Different word *n*-gram models have sometimes been used for different speakers (Louwerse & Crossley, 2006; Sridhar et al., 2009). In both these studies, the models were evaluated on the Map Task corpus where the instruction giver and instruction follower might use words differently. In (other) machine learning and deep learning approaches, speaker cues have typically been encoded using a feature that tells the model who the speaker is (Bothe, Weber et al., [2018]; Cerisara et al., 2018; Di Eugenio et al., 2010) or whether the previous utterance is uttered by the same or different speaker from the current utterance (e.g., Lager & Zinovjeva, 1999; Liu et al., 2017; Yano et al., 2021; Zhao & Kawahara, 2019). Interestingly, many deep learning models, despite encoding contextual cues such as the structure of turns directly into the model, do not take into account who uttered which utterance, and still achieve a good performance. This is rather surprising as it

goes directly against the theoretical claims that speech acts can only be considered in their interactional context (Clark, 1996; Schegloff & Sachs, 1973).

For the same reasons, it might be surprising that turn and discourse organization cues have seldomly been used in dialog act classification studies. In fact, information on the location of the utterance in the turn can indeed be beneficial for dialog act classification (Yano et al., 2021). Finally, specific to the Map Task corpus, Di Eugenio et al. (2010) have annotated dialogs with a label denoting the subdialog they are part of.

2.4. Interim conclusion

The overview of the different categories in the different studies together with the performance of the dialog act classification provides an insight into the dialog act classification systems and the linguistic cues that have been used over the years.

The literature overview shows that the studies may vary widely in the approach taken (i.e., frequency-based, machine learning, and deep learning) and the cues utilized, but that differences in performance on each corpus across the dialog act classification systems fall within a small 10%–15% accuracy window. Due to the differences in setup, corpora, approaches, and models used, it is difficult to draw more definitive conclusions on which cues are most important in dialog act classification. This is complicated further by the fact that the machine learning and deep learning approaches, given the very nature, do not allow for insights in what linguistic cues contribute to the performance of the models. Moreover, while the frequency-based approaches do provide some insight into the cues that contribute to dialog act classification, only a fraction of all cues has been investigated in such approaches.

In short, from the literature overview given here, we can conclude that (1) most dialog act classification studies use machine learning and deep learning approaches, which are inherently nontransparent, (2) the differences between the transparent frequency approaches and the machine learning and deep learning approaches are small, (3) none of the studies included all linguistic and contextual features identified in the literature making it difficult to determine their individual contribution to dialog act classification performance, and (4) exacerbated by the previous three points, given that none of the studies use all three corpora (Switchboard corpus, MRDA corpus, and Map Task corpus), it is unclear how the contribution of the features to dialog act classification differ or generalize across the three corpora.

3. Analyzing the role of surface and contextual linguistic cues

Instead of focusing on maximizing the performance, as is the goal in the dialog act classification, the aim of the current paper was to shed light on the role of different surface and contextual linguistic cues in the identification of the dialog act. We, therefore, focused on the causal relationships between the independent variables (linguistic cues) and the dependent variable (dialog act), asking the question to what extent each individual cue can explain the dialog act. To this end, we (1) focused not just on one corpus but on three, each being very different in the design of the corpus and dialog act tagset and in the type of dialog,

(2) used an explanatory paradigm instead of a predictive one, to increase the transparency and interpretability of the findings on the informativeness of the different cues (Jones, 2017; McClelland, 2009), (3) treated these cues individually and together in a very systematic way, and (4) compared the findings across the three corpora. This does not only allow us to compare categories, but also determine the importance of each category and cue for explaining the dialog act and allow us to interpret these findings more generally.

One of the common reasons to prefer a prediction-based approach over an explanation-based one is that the latter is more prone to overfitting and, therefore, does not typically generalize as well as prediction-based models (Yarkoni & Westfall, 2017). A lack of generalization should, however, not be an issue in the current study. First, we are investigating (especially from a cognitive science perspective) large corpora. Moreover, we investigate different corpora, allowing us to compare and interpret the findings in a broader scope. Finally, we are comparing linguistic cues on the same model, where the actual fit is thus interpreted relative to other fits, rather than to a baseline (as is common in computational linguistics).

It is important to mention that explanatory power should not be directly associated with predictive power, due to the different goals of the models in each of the paradigms (Douglas, 2009; Rosenberg et al., 2018; Yarkoni & Westfall, 2017). Yet, specific models used for prediction can also be used for explanation. Logistic regression is one such model. It is commonly used as a predictive modeling technique, for instance, by Ang et al. (2005) and Sridhar et al. (2009) in dialog act classification, but can also be used as an explanatory or interpretable model providing insights into the relationships between the input features and the outcome variable.

3.1. Methods

3.1.1. Corpora

In an empirical evaluation of the role of surface and contextual linguistic cues in dialog act classification, we used the same three commonly used corpora as in the overview (Table 1): the Switchboard corpus, the MRDA corpus, and the Map Task corpus. Besides being among the most-used English language corpora in dialog act classification, these corpora were specifically selected because they differ considerably from each other on multiple dimensions: (1) the conversational setting, (2) the size of the corpus, and (3) the dialog act set with which the corpora were annotated. These dimensions are summarized in Table 3.

The Switchboard corpus contains spoken telephone conversations on a variety of self-selected topics between two strangers (Godfrey, Holliman, & McDaniel, 1992). A subset of these dialogs was annotated with dialog acts, and subsequently, referred to as the Switchboard Dialog Act corpus (Jurafsky et al., 1997). With 1,155 conversations annotated with dialog acts, it is the largest of the three corpora included in our computational analysis. Originally, the corpus was annotated with a set of 220 multi-dimensional Switchboard-DAMSL dialog acts, but which was later clustered into a set of 42 dialog acts (Jurafsky et al., 1997). In our analysis, we concatenated those utterances that were broken into two parts because of a noninterruptive backchannel, as these utterances ought to be seen as one. We also removed

Table 3
An overview of the properties and statistics of the corpora

Corpus	Task-based	Strict roles	Participants could see each other	Participants per dialog	Dialog acts	Dialogs	Utterances	Average utterances per dialog	Types	Tokens	Average tokens per sentence
Switchboard	−	−	−	2	41	1,155	201,182	174.2	21,784	1,474,011	7.33
MRDA	±	±	+	4–10	5	73	108,963	1492.6	12,599	762,736	7.00
Map Task	+	+	±	2	13	128	27,084	211.6	2333	153,780	5.68

Note. "+" and "−" refer to the presence and absence of that dimension in the corpus, respectively. When a corpus consists of conversations where a dimension is less well-defined, it is indicated by "±."

any nonverbal tags, resulting in a total of 41 different dialog acts. After all, nonverbal tags do not contain any surface information.

The MRDA corpus contains dialogs that were recorded at meetings at the International Computer Science Institute (Janin et al., 2003). The meeting topics were related to the creation of the corpus itself. Unlike the other corpora that included dyads, the number of participants ranged between 4 and 10. The corpus consists of 75 dialogs and, like the Switchboard corpus, the MRDA corpus was also annotated with a multi-dimensional set of 55 dialog act labels (Shriberg et al., 2004). A major difference in annotation compared with the Switchboard corpus, however, is that one utterance could have more than one label associated with it. Each utterance was annotated with one of the 11 general labels and could contain multiple optional specific labels or disruption forms. In dialog act classification, these dialog acts are, however, generally simplified into a smaller set, of which a set of five (one-dimensional) dialog acts is used most often (Ang et al., 2005; Lee & Dernoncourt, 2016).

The HCRC Map Task corpus contains dialogs in a task-based setting, where the goal of the task was to solve a road map task collaboratively (Anderson et al., 1991). An instruction giver guided an instruction follower through a road map, where both participants received a slightly different version of the map. The instruction giver and follower could not see each other's map, so an interaction was necessary to solve the task. The Map Task corpus contains 128 dialogs, of which half were dialogs between participants who could not see each other. Each participant, furthermore, participated in two dialogs. The corpus was annotated with 13 different dialog acts.

We automatically annotated Switchboard and the MRDA corpus with PoS tags using the Stanford Neural Pipeline (Qi, Zhang, Zhang, Bolton, & Manning, 2020). This is a state-of-the-art PoS tagger that was trained on the Universal Dependencies tagset, which has a generic tagset and is language-independent (Petrov, Das, & McDonald, 2012). The Universal Dependencies tagset consists of 17 tags. Because the Map Task corpus has already been manually annotated with a set of 58 PoS tags, we used these tags in the analyses.

As can be observed in Table 3, the three corpora differ in the total number of dialogs as well as in the average number of utterances for each dialog. The Map Task corpus is about a fifth of the size of the MRDA corpus, and the MRDA corpus is about half the size of Switchboard, yet all three corpora contain a large number of observations (utterances). There are also considerable differences in the number of utterances per dialog with the MRDA corpus containing dialogs that are about seven times longer than the other corpora. In addition, utterances are significantly shorter in the Map Task corpus. The three corpora also differ in what kind of dialog they contain. While the Map Task corpus contains dialog that revolves around a clearly defined task, the conversations in the Switchboard were completely free-flow with no clearly defined goal. Similarly, participants in the Map Task corpus had a clearly defined role with one being the instruction giver and the other one being the instruction follower, but there were no such speaker roles in the dialogs of the Switchboard corpus. In terms of dialog and speaker roles, the MRDA corpus operates between Switchboard and Map Task corpus: Even though there was no clearly defined goal for the meetings that are recorded, given the nature of the meeting, the dialog was not completely free-flow either, and similarly, while each participant in the MRDA corpus had a specific expertise, this did not necessarily restrict their linguistic

Table 4
An overview of the 10 most frequent dialog acts

Switchboard		MRDA		Map Task	
Dialog Act	%	Dialog Act	%	Dialog Act	%
Statement-nonopinion	37.48	Statement	58.95	Acknowledge	20.69
Acknowledge (backchannel)	19.08	Disruption	14.02	Instruct	15.75
Statement-opinion	13.19	Backchannel	13.42	Reply-yes	11.93
Uninterpretable, abandoned or turn-exit	7.79	Floor grabber	7.20	Explain	7.98
Agree/accept	5.55	Question	6.41	Check	7.89
Appreciation	2.38			Ready	7.61
Yes-no-question	2.36			Align	6.56
Yes answer	1.51			Query-yes-no	6.49
Conventional-closing	1.28			Clarify	4.40
Wh-question	0.99			Reply-wh	3.38

Note. The tagset of the MRDA corpus consists of only five dialog acts.

freedom in a similar way as in Map Task corpus. In addition, the corpora differed with regard to the availability of visual information. In the Switchboard corpus and in half of the Map Task corpus, the dialog participants could not see each other and thus also could not potentially leverage visual cues for the identification of the dialog act (cf. Bucciarelli et al., 2003; Domaneschi et al., 2017; Enrici et al., 2011; Jang et al., 2014). Finally, and importantly, the size of the set of dialog acts varied considerably between the corpora with the MRDA corpus consisting only of five dialog acts, the Map Task corpus 13, and the Switchboard 41. These differences could be seen as a drawback when comparing the corpora. In the current study, however, these differences allow for investigating similarities in dialog act classification.

Table 4 gives an overview of the most frequent dialog acts in each corpus. What the three corpora have in common is that the frequency distribution of dialog acts is highly skewed. Interestingly, the degree of skewedness roughly correlates with the performances reached on the three corpora by the dialog act classification studies in the literature, with performances generally being the highest on the MRDA corpus and lowest on the Map Task corpus (see Table 2). While this is primarily a concern for prediction models, it is still an important consideration when trying to understand the mechanisms behind speech act recognition. This skewedness might in fact be an important characteristic of dialog acts, and more broadly, speaker intentions. At other levels of communication, such as words, it seems to be beneficial to have more skewed distributions for both language learning (Hendrickson & Perfors, 2019) and language use (Ferrer-i-Cancho, 2018; Zipf, 1949). Skewed distributions are also widely present in cognitive systems and could be the result of adaptive processes of the language system (Kello et al., 2010; Linders & Louwerse, 2023).

There also seems to be a large degree of arbitrariness in the choice of dialog acts and their associated frequencies, as it is virtually impossible to map each set to the set of dialog acts of another corpus. Moreover, there is a large degree of variation in the frequencies of dialog acts that do (partially) overlap, such as the statement-like "Instruct" only occurring less than 16%, while the two statement dialog acts of the Switchboard corpus combined and the statement dialog act of the MRDA corpus have a frequency of above 50%. Similarly, 12% of the dialog acts are yes answers in the Map Task corpus, while they are less than 2% in the Switchboard corpus. These differences are likely due to differences in the type of dialog, the setup, and the annotations. These differences do not imply, however, that a direct comparison of the cues on the three corpora is not warranted, because the underlying mechanisms are still likely the same.

3.1.2. Computational implementation

In order to investigate the explanatory power of the different linguistic cues, we used a multinomial logistic regression model. The model was, unlike in dialog act classification, trained on the whole corpus. Logistic regression is a conditional model that tries to minimize the cross entropy of a model with features $F = [f_1, f_2, \ldots, f_l]$ and a set of weights for each feature (as indicated by the number) and dialog act combination (as indicated by d) $W_d = [w_1^{(d)}, w_2^{(d)}, \ldots, w_l^{(d)}]$, where the weights for each dialog act sum to 1: $\sum_{i=1}^{l} w_i = 1$. The probability of a dialog act d can then be calculated using a simple linear predictor function, which can be represented by the dot product between a vector with feature probabilities and weight vector for each dialog act,

$$P(d|F; W_d) = \sum_{i=1}^{l} P(f_i|d)\, w_i^{(d)} \tag{1}$$

Logistic regression then uses the softmax function to find the dialog act d with the largest probability:

$$d = argmax_{d'} \frac{e^{P(F|d')\cdot w^{(d')}}}{\sum_{d''} e^{P(F|d'')\cdot w^{(d'')}}} \tag{2}$$

The denominator is the sum over all dialog acts d'' and serves as a normalization constant to turn the weights into a probability.

We preprocessed each corpus by removing punctuation marks and annotation-related comments, as they can have a significant impact on the performance (Duran et al., 2023; Ribeiro et al., 2019a). We converted all characters to lower-case and removed dialog acts with a nonverbal tag in Switchboard and the unlabeled utterances in the MRDA corpus. This affected roughly 2% of the data for Switchboard and MRDA. Data of the Map Task Corpus were unaffected. The discarded data were excluded from the statistics reported in Table 3.

We encoded the cues from Fig. 1 into a feature matrix encoding. Each individual feature vector was then treated as a separate independent variable in the regression model. We used primarily three different techniques for the encoding of the cues. Categorical cues, such as

n-grams, were converted to a one-hot vector encoding where each dimension represents a single *n*-gram. In the current study, we only investigated unigrams of different linguistic units, such that we are only measuring the impact of the unit information and are not conflating structural information in the units as well. We created character, word, and PoS unigrams and word unigrams of the previous utterances for the contextual representation of surface information. Unigrams with a frequency of one are not very useful to the model. To enhance their usefulness, we created a catch-all group for all words with a frequency of one. Next to a representation of context through surface information, the dialog acts themselves contain valuable cues. Just like the dialog acts in Table 1, we, therefore, also included the last dialog act as a contextual cue, modeled using a one-hot encoding, thus treating each dialog act as an independent binary variable.

Dense distributional semantic cues were encoded as ordinal cues with each dimension being treated as a single ordinal variable. For the noncontextualized word embeddings, pretrained Global Vectors (GloVe) for word representation were used to map all words to a 300-dimensional semantic vector (Pennington, Socher, & Manning, 2014). GloVe has often been used in deep learning models of dialog act classification (e.g., Chen et al., 2018; Lee & Dernoncourt, 2016; Raheja & Tetreault, 2019; Wan et al., 2018). All words for which vectors were available were vectorized. The word vectors were then averaged to obtain a sentence representation. Each dimension was treated as a separate independent variable in the model. For the contextual word embeddings, we used 768-dimensional sentence vectors that were created using Sentence Embeddings, which is a framework used to extract meaningful sentence embeddings from transformer models (Reimers & Gurevych, 2019). We used MPNet for extracting the sentence vectors, since MPNet was specifically trained on sentence similarity, and since to our knowledge, it currently reaches the highest performance on tasks involving sentence similarity (Song, Tan, Qin, Lu, & Liu, 2020). The added benefit of contextualized word embeddings is that they encode the whole utterance at once, thereby also encoding the structural factors that make up an utterance. But this come at the cost of interpretability: It is impossible to discern the different factors that are expressed in the resulting vector representations. For example, it is likely that contextualized word embeddings encode not just lexical information, but also other information, such as the sentence structure (Hu, Gauthier, Qian, & Levy, 2020; Warstadt et al., 2020).

Finally, we included several cues that were treated as ordinal variables. Similar to Zhao and Kawahara (2019), we encoded the speaker of the last dialog act using a binary encoding, modeling whether the utterance was uttered by the same speaker as the current utterance or by a different one. Finally, the utterance length and utterance position cues were encoded in a one-dimensional ordinal vector.

We included all the surface and contextual linguistic cues outlined by the literature and illustrated in Fig. 1 with the exception of the latent semantic analysis and the discourse organizational cues. Since latent semantic analysis is a different approach to modeling dialog act cues that is incompatible with the current implementation of logistic regression, it was discarded. The discourse organizational cues are specific to the Map Task corpus and unavailable for the other corpora, and were, therefore, also discarded.

Table 5
Performances of logistic regression with a single cue

	Switchboard (201,182)			MRDA (108,963)			Map Task (27,084)		
Cue	χ^2	IVs	R^2	χ^2	IVs	R^2	χ^2	IVs	R^2
Character unigrams	346,797	34	.406	96,692	34	.363	45,113	31	.355
Word unigrams	590,547	13,169	**.692**	144,032	7563	**.541**	79,960	1379	**.630**
PoS unigrams	295,273	17	.346	76,449	17	.287	44,758	57	.353
Noncontextualized word embeddings	441,983	300	**.518**	113,822	300	**.427**	61,948	300	**.488**
Contextualized word embeddings	559,483	768	**.655**	146,448	768	**.550**	79,808	768	**.629**
Utterance length	225,008	1	.264	54,903	1	.206	19,721	1	.155
Previous dialog act	111,575	42	.131	1712	6	.006	20,656	14	.163
Surface cues of context	235,040	13,457	.275	36,341	7750	.136	29,780	1433	.235
Previous speaker	43,232	1	.051	6583	1	.025	4298	1	.034
Utterance position	37,953	1	.044	6891	1	.026	4428	1	.035

Note. The three cues yielding highest explained variance per corpus are marked in bold. The total number of observations (i.e., utterances) per corpus are summarized next to the corpus name.

3.2. Results

3.2.1. Individual cues

We first analyzed the explanatory power of the cues in isolation. To understand the individual contribution of each cue, we ideally would have preferred to report the explained variance of the model with a particular set of variables in the relationship with dialog acts. However, for logistic regression, no explained variance can be computed reliably (Mittlböck & Schemper, 1996). Instead, we reported the goodness of fit, equivalent to the deviance of the residuals of each regression model using χ^2. The residual deviance was calculated by subtracting the deviance of the regression model from the deviance of a reference (null) model with no cues and only an intercept. Moreover, from these two deviances, we derived an approximate of the actual explained variance: the McFadden pseudo R^2 coefficient (McFadden, 1974), using the following formula:

$$\text{McFadden Pseudo } R^2 = 1 - \frac{\text{Deviance}_{\text{model}}}{\text{Deviance}_{\text{null}}} \tag{3}$$

The McFadden pseudo R^2 coefficient is an approximation of the actual explained variance. Nonetheless, this coefficient seems suitable for such analyses (Azen & Traxel, 2009), and can, moreover, be intuitively determined from comparing the deviances of the model with the predictors and the reference model.

The results for the Switchboard, MRDA, and Map Task corpora are summarized in Table 5. There are two major observations that can be drawn from these results. First, the trends of the explained variance of the different cues are remarkably similar across corpora, despite differences in the type of dialog and dialog act tagsets. Second, the cues with the most explanatory

power are the same across corpora: contextualized word embeddings, word unigrams, and noncontextualized word embeddings, showing the prominence of lexical-level cues.

The fact that trends are similar across corpora, despite the very different types of dialog and tagsets, shows that the underlying statistical patterns are remarkably similar. This allows us to generalize the findings about the statistical behavior of dialog acts across these dimensions where the corpora differ. While according to the literature overview, predicting dialog acts seems to be the easiest on the MRDA corpus and the most difficult on the Map Task corpus, this is not visible in the explanatory power of the cues. Interestingly, the skewedness of the dialog act frequencies does seem to correlate with the prediction accuracies across corpora, suggesting that the differences in prediction accuracy observed in the literature across corpora is primarily caused by the skewedness of the frequency distribution.

Word unigrams surpass contextualized word embeddings on both the Switchboard and Map Task corpora. This is very surprising given that contextual word embeddings use additional information in the pretraining phase, unlike word unigrams and noncontextualized word embeddings. Contextual word embeddings also likely encode structural information and provide a more complex and dense representation of the utterance. Noncontextualized word embeddings come in third on all three corpora. This prominence of word-level encodings, and in particular, word frequency statistics is consistent with the findings on predicting dialog acts (Cerisara et al., 2018; Duran & Battle, 2018), although these studies encoded the words of an utterance in a deep neural network, which includes the structure of the utterance, making it difficult to conclude the exclusive role for word unigrams. We encoded word unigrams exclusively, and as such, the utterance structure is completely absent.

Surface cues seem overall to be more indicative of the dialog act than contextual cues, especially on the MRDA corpus. This is exemplified through one of the simplest cues, the utterance length, which gives a higher performance than all contextual cues on the MRDA corpus. Importantly, however, perhaps the prominence of surface cues can be explained by the fact that the model is considering only one cue at the same time. And perhaps contextual cues are in fact very important in a model that contains both surface and contextual cues. At the same time, it is likely that, given the flexibility of dialog even in rather constrained contexts, such as in the Map Task corpus, it is virtually impossible to explain the dialog act just from the context. And just like all the dialog act classification studies use surface cues, surface cues form the backbone of speech act recognition. This would mean that rather than forming important units in dialog, adjacency pairs (Schegloff & Sachs, 1973) form the exception rather than the rule. To investigate this hypothesis, we consider a regression model that includes different subsets of cues.

3.2.2. Cue combinations

Computing the explanatory power of each cue individually by training the logistic regression model on a single cue at the same time gives us a good indication of the importance of each cue in dialog act identification in isolation. However, it does not represent the actual importance of cues in human speech act recognition where likely multiple cues are competing with each other. Moreover, it does not take into account potential multicollinearities between different cues. For example, character and word unigrams likely capture similar information,

but are defined at different linguistic levels. We, therefore, also explored the relative importance of each cue through a hierarchical logistic regression analysis, where we progressively added cues to the model and investigated their added value.

Standardized regression coefficients (beta coefficients) are commonly computed to investigate the importance of the different cues in a regression analysis. However, in our case, a single cue can consist of multiple independent variables. For example, each unigram constitutes a single independent variable. Hence, we have a beta coefficient for each independent variable, rather than for each cue. Consequently, these beta coefficients do not directly provide the answer to which cues are most important. Moreover, due to the possible multicollinearities between different cues, such beta values might become unreliable (Alin, 2010). Instead, we opted for a hierarchical logistic regression, investigating the performance of multiple models trained on different sets of cues separately, rather than investigating a single model trained on all cues.

Because word unigrams and contextualized and noncontextualized word embeddings all target semantic information in the utterance at a word level, and because contextualized and noncontextualized word embeddings are less transparent and perform on and below par compared to word unigrams, respectively, we excluded both embeddings from the hierarchical regression analysis. However, for completeness, we trained a model that included all cues, including the contextualized and noncontextualized word embeddings.

The hierarchical logistic regression procedure was implemented as follows. At each step, we added a new cue to the model, then retrained the model, evaluated its explanatory power, and finally, compared the change in model performance to the previous cue combination. We entered the linguistic cues according to the order in Fig. 1, thus starting with surface cues. The surface cues have been ordered on their linguistic level, starting with the lowest, that is, the characters. The contextual cues were ordered according to their (hypothesized) concreteness in capturing the context. That is, the previous dialog acts provide the most direct cue for the next dialog act, as exemplified using, for example, adjacency pairs (Schegloff & Sachs, 1973), while utterance position provides the least concrete contextual information. However, to avoid that findings might be biased by the order in which the linguistic cues were entered, we also reversed the ordering, starting with the contextual cues, and in particular, with the utterance position cue.

The performances of different cue combinations in the hierarchical logistic are summarized in Table 6 for the subsets starting with surface cues, and in Table 7 for the subsets starting with contextual cues. For completeness, we also evaluated a model with all cues, including the two embedding cues. First note that again the trends across corpora are very similar. There are two observations that we can draw from this analysis: (1) the word level again has a high prominence, both as a surface cue (word unigrams), and as surface information of the context, and is not outcompeted by any other cue, and (2) the surface and contextual cues complement each other, while the different cues within each category mostly do not.

Adding word-level cues has the largest effect on the (approximated) explained variance. This is true for both the word unigrams and the surface information of the context, that is, word unigrams of the previous utterance, and is illustrated in Tables 6 and 7 as being the cues with the highest change in explained variance when added to the cue set. In line with the

Table 6
Performances of logistic regression with a combinations of cues starting with surface cues

	Switchboard (201,182)				MRDA (108,963)				Map Task (27,084)			
Cue	χ^2	IVs	R^2	ΔR^2	χ^2	IVs	R^2	ΔR^2	χ^2	IVs	R^2	ΔR^2
Character unigrams	346,797	34	.406		96,692	34	.363		45,113	31	.355	
+ Word unigrams	590,207	13,203	.691	**.285**	144,526	7597	.543	**.180**	80,395	1410	.633	**.278**
+ PoS unigrams	592,797	13,220	.695	.003	146,005	7614	.548	**.006**	81,955	1467	.646	.012
+ Utterance length	591,109	13,221	.693	−.002	145,979	7615	.548	.000	81,907	1468	.645	.000
+ Previous dialog act	648,677	13,263	.760	**.067**	147,493	7621	.554	.006	92,037	1482	.725	**.080**
+ Surface cues of context	727,408	26,720	.852	**.092**	174,583	15,371	.656	**.102**	103,999	2915	.819	**.094**
+ Previous speaker	729,887	26,721	.855	.003	174,993	15,372	.657	.000	104,265	2916	.822	.002
+ Utterance position	728,534	26,722	.854	−.002	175,111	15,373	.658	.001	104,638	2917	.824	.003
+ noncontextualized and contextualized word embeddings (all cues)	763,146	27,790	.894	.041	198,198	16,441	.744	.087	119,358	3985	.940	.116

Note. The three cues yielding highest change in explained variance per corpus are marked in bold.

Table 7
Performances of logistic regression with a combinations of cues starting with contextual cues

	Switchboard (201,182)				MRDA (108,963)				Map Task (27,084)			
Cue	χ^2	IVs	R^2	ΔR^2	χ^2	IVs	R^2	ΔR^2	χ^2	IVs	R^2	ΔR^2
Utterance position	37,953	1	.044		6891	1	.026		4428	1	.035	
+ Previous speaker	47,974	2	.056	.012	7509	2	.028	.002	4772	2	.038	.003
+ Surface cues of context	270,750	13,459	.317	**.261**	43,065	7752	.162	**.134**	32,731	1435	.258	**.220**
+ Previous dialog act	319,491	13,501	.374	.057	44,162	7758	.166	.004	41,125	1449	.324	.066
+ Utterance length	487,894	13,502	.572	**.198**	93,325	7759	.350	**.185**	55,675	1450	.439	.115
+ PoS unigrams	534,821	13,519	.627	.055	111,774	7776	.420	.069	76,950	1507	.606	**.168**
+ Word unigrams	731,751	26,688	.857	**.230**	174,654	108,963	.656	**.236**	104,424	2886	.823	**.216**
+ Character unigrams	729,478	26,722	.855	−.002	175,063	15,373	.657	.002	104,650	2917	.825	.002
+ noncontextualized and contextualized word embeddings (all cues)	763,146	27,790	.894	.041	198,198	16,441	.744	.087	119,358	3985	.940	.116

Note. The three cues yielding highest change in explained variance per corpus are marked in bold.

findings in the literature overview presented earlier, this again highlights the important role of lexical frequency statistics in determining the dialog act.

Table 6 shows that adding any additional surface cues to the combination of character and word unigrams has relatively little effect on the model's performance. However, adding contextual cues to the model does in fact significantly increase the explanatory power of the model. In a somewhat similar fashion, Table 7 shows that combining different contextual cues has little effect on the model's performance, except for the surface cues of the context. Adding a surface linguistic cue to the model, however, significantly increases the explanatory power of the model. This shows that while the different surface and contextual cues within each respective category are mostly not complementary, the two categories themselves are.

Yet, the effects are not the same when comparing both orders in which the cues were added to the model. The effect of adding contextual cues to a model with surface cues is comparatively smaller than the effect of adding surface cues to a model with contextual cues. Surface linguistic cues, therefore, seem the predominant cues for classifying a dialog act, while contextual cues can complement these surface cues.

There are, furthermore, some notable differences between the corpora. While we saw a smaller role for contextual cues in isolation, this effect is less pronounced when multiple cues are added to the model. Yet, the previous dialog act seems to have a very low explanatory power across both analyses. Moreover, we can observe that PoS unigrams play a slightly larger role in the Map Task corpus, likely due to the use of a larger and more detailed set of PoS tags.

Combining all cues leads to an explained variance approximation of 66%–86% without the embeddings included and to an explained variance approximation of 74%–94% with these embeddings included. Interestingly, the contextualized and noncontextualized do explain an additional 4%–12%, and thus do provide complementary information compared with all other cues. However, the goal of the current paper was not to optimize performance but to understand the relative importance of surface and contextual linguistic cues.

The R^2 values reported here ought to be considered with caution because they are estimated. However, despite the fact that they ought to be treated with caution, we might want to speculate about the reasons why variance of 26% to 6% remains unaccounted for. First, there may be other cues that are important in the identification of the dialog act that have not yet been utilized. In particular, for the MRDA corpus, extralinguistic cues might be important, because the participants could utilize visual information (see Table 3). And second, there might also be significant variance in the realization of a dialog act between different speakers or in different contexts, making it difficult to establish an exact mapping between the cues and the dialog acts in all cases.

Finally, there is a marginal influence of utterance position and speaker cues in both analyses. This corresponds to findings in the dialog act classification literature, where marginal increases of up to 2% were found when speaker cues (Zhao & Kawahara, 2019), utterance position cues (Yano et al., 2021), and a combination of speaker, utterance position, and utterance length cues (Li & Wu, 2016) were added to a baseline of surface linguistic cues. In fact, as we observed in the literature overview, many deep learning approaches achieve a high performance without considering the structure of turns, nor who uttered which utterance. But

while these findings are perhaps unsurprising in light of the computational task of dialog act classification, given their small vector sizes, the findings are rather surprising from a cognitive science perspective, because they suggest that human speech act recognition does not rely on where in the turn the dialog act occurs, nor whether the speaker of the previous dialog act is the same as the current one. This seems to go against the arguments that speech acts can only be considered in their interactional context (Clark, 1996; Schegloff & Sachs, 1973). However, it is important to consider that we have treated these cues as independent variables in our model. Perhaps, if these cues are considered jointly with other cues, such as word *n*-grams, they might prove useful (see, e.g., Louwerse & Crossley, 2006).

4. General discussion

The current study discussed the role of surface linguistic and contextual linguistic cues in the identification of speech acts. Because of the wealth of studies in the computational field of dialog act classification and the potential of these studies to inform cognitive science on the mechanisms behind speech act recognition, we investigated the literature on dialog act classification. We identified different linguistic cues and their potential to identify dialog acts through a review of 50 dialog act classification studies, and classified the 15 different cues in these studies into surface linguistic and contextual linguistic cues. The surface linguistic cues included different linguistic levels, from character, word, syntactic to semantic levels. The contextual linguistic cues included the previous dialog acts, surface cues from the previous utterances, speaker cues, and utterance position cues. The reviewed studies used frequency-based, machine learning, and deep learning models to classify dialog acts. When comparing the performance, we found that more complex models (e.g., deep learning > machine learning > frequency-based) achieved higher performance; however, differences were small. Moreover, there were no indications which specific surface or contextual cues best explained optimal performance for dialog act classification. We concluded that this could be due to the differences in setup, corpora, and combinations of cues across the different studies, and therefore, aimed to shed light on the importance of the 15 linguistic cues ourselves by systematically investigating their explanatory power both individually and combined. We aimed at comparing the explanatory across corpora to shed light on the behavior of the cues in different settings.

We observed very similar trends in explanatory power across the different corpora. We saw a dominance of surface cues over contextual cues. Zooming in on the individual cues, we, furthermore, saw a dominance of cues quantifying information at the lexical level. In particular, word unigrams scored on par with or even better than the more complex contextual word embeddings, suggesting that word frequency statistics is very important in dialog act classification (Duran & Battle, 2018; Louwerse & Crossley, 2006). It is, however, important to note that the construction of the dialog act tagsets and the annotation seemed to have relied significantly on cues from the words and the word sequences in utterances, since the annotations were based primarily on transcriptions (Jurafsky et al., 1997). The question remains whether this has biased the performances of our models that contained these cues.

Nevertheless, the observation that word frequency statistics is important in dialog classification is not new. An encoding that leverages these frequency statistics of words has outperformed encodings based on noncontextual word embeddings (Cerisara et al., 2018; Duran & Battle, 2018), suggesting that these statistics are more important than the semantics of words. However, one major difference between these approaches and our approach is that these studies used a deep learning approach, thereby also leveraging word order information. In our approach, we drew the same observations from analyzing word unigrams alone (thus not conflating any structural information in the cue) in an explanatory model. Perhaps both contextualized and noncontextualized word embeddings are not the most optimal cue in speech act recognition. Noble and Maraev (2021) found contextualized word embeddings to be suboptimal for dialog act classification, as the performance without fine-tuning on dialog corpora or specifically on dialog act classification was rather poor. They, however, argued that, contrary to the cue themselves not being optimal, the data that the pretrained embeddings were trained on are not optimal because written language is very different from the spoken dialog (cf. Biber, 1988; Clark, 1996; Louwerse, McCarthy, McNamara, & Graesser, 2004).

The prevalence of surface linguistic cues and of word frequency statistics, in particular, furthermore supports the argument that illocutionary force indicating devices play an important role in identifying the speech act of an utterance, since these are primarily based on the surface linguistic information. The interactional function of speech acts, such as defined in adjacency pairs (Schegloff & Sachs, 1973) or as joint acts (Clark, 1996), showed to be less important than the lexical cues in both analyses.

There were also some notable differences between corpora, which help shed light on mechanisms behind the identification of dialog acts, something that we could only observe due to the comparisons across corpora. While the role of surface linguistic cues seemed consistent across corpora, the role of contextual cues differed across the corpora. The role of context was the smallest for the MRDA corpus, that corpus consisting of meeting dialogs. One possible reason for this is that these dialogs contain more (4–10) participants, which might result in more flexibility in the dialog structure as more participants can chime in and direct the conversation. An alternative explanation is that because there are only five dialog acts in the MRDA corpus, these distinctions between dialog acts are not very fine-grained and hence do not require any contextual disambiguation.

In this paper, we have approached dialog act classification as not just a challenge for a single corpus with a specific type of dialog or specific tagset, but from a cross-corpus perspective relying on a principle of parsimony. This allowed us to more broadly investigate the role different linguistic cues play and allowed us to make generalizations. Here, we deviated from the existing computational linguistics literature, which typically focuses on individual corpora and achieving the highest performance. Results were mostly consistent across corpora, with similar trends across corpora when compared with a majority vote baseline and with the general performance correlating with the skewedness of the frequency distribution of the dialog acts. Such kind of generalizations across corpora can in turn lead to the validation of different tagsets as a proxy for the way we humans categorize speaker intentions and eventually to a formal framework of speech acts. Similarly, by explaining differences in

performance between corpora through differences in the nature of the dialogs or in the tagset, can also be important for such a validation.

The current paper did not take the common computational linguistic approach of finding the model and cue combination that leads to the highest performance of dialog act classification. Instead, we used a cognitive science approach by focusing on the explanatory power with the goal of investigating the underlying mechanisms (cf. Jones, 2017; McClelland, 2009), applying the principle of parsimony (cf. Chater & Vitányi, 2003). We thereby aimed to bridge the gap between the computational linguistic work on dialog act classification — typically not taking into account mechanisms of human cognition — with the cognitive science work on speech acts — typically not focusing on the categorization of speech acts and the role of different linguistic cues in their identification. The wealth of studies on dialog act classification is able to provide important insights into the mechanisms behind speech acts and, vice versa, a cognitive science perspective on the dialog act classification literature can provide insights into the role of different approaches and cues on the performance of computational models of dialog act classification.

Since Austin (1962) asked the question how to do things with words over five decades ago, speech acts have received considerable attention in the cognitive science literature, from studies in philosophy (Searle, 1976), linguistics (Levinson, 1983; Wierzbicka, 1987), psychology (Clark, 1996), to artificial intelligence (Jurafsky & Martin, 2009). Over the last two decades, the attention for speech acts has shifted toward computer science with a focus on predicting in an interactional context. That shift can be explained by the advances in natural language processing, combined with an interest in the use of mixed-initiative dialogue in conversational systems (Hearst, 1999), including intelligent tutoring systems (Graesser et al., 2004).

Most recent studies have focused on machine learning and deep learning techniques to predict the dialog act in a corpus, particularly, the corpora central to the current study, the Switchboard corpus, the MRDA corpus, and the Map Task corpus. This is useful, for instance, to better understand the intention of speakers in conversational systems. However, the machine learning and deep learning approaches to optimize the performance of dialog act classification have obfuscated the features explaining the identification of the dialog act.

The current study has shown that in the features considered by machine learning and deep learning studies to predict dialog acts, surface linguistic features played a role. In our own analyses, we have shown that surface linguistic features played the most prominent role in explaining dialog act classification. It is ironic that the evidence for surface linguistic cues we find in a frequency-based approach is what research on speech act classification started out with several decades ago, the illocutionary force indicating devices (IFIDs) — explicit linguistic cues.

Elsewhere, we have emphasized the importance of the role of linguistic cues in language, explaining cognitive effort (Linders & Louwerse, 2023) and perceptual information (Louwerse, 2018, 2021). The conclusions of the current study are in a similar vein. Even though dialog acts can be explained by different aspects of human communication, nonverbal and verbal cues, contextual and surface linguistic cues, when applying the principle of parsimony, lexical information best explains dialog act classification, a finding relevant for all aspects of cognitive science — computational and cognitive.

Acknowledgments

This research has been funded by a grant PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province of Noord-Brabant, and the municipality of Tilburg awarded to the second author. The usual exculpations apply.

Notes

1 We will use the term "speech acts" to refer to illocutionary acts in general. The term "dialog acts" is used as a specific subset of speech acts, reserved for dialog.
2 The 80.1% accuracy that the authors reported on Switchboard could not be replicated by Kumar, Agarwal, Dasgupta, and Joshi (2018), and Papalampidi, Iosif, and Potamianos (2017). Kumar et al. (2018) reported that personal correspondence with the authors revealed they had used a nonstandard test set. The accuracy reported here was taken from Kumar et al. (2018).
3 It is unclear whether the authors used a feature encoding, based on characters or words.

References

Abbeduto, L., Furman, L., & Davies, B. (1989). Identifying speech acts from contextual and linguistic information. *Language and Speech*, *32*(3), 189–203. https://doi.org/10.1177/002383098903200301

Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019). Contextual dialogue act classification for open-domain conversational agents. In B. Piwowarski, M. Chevalier, & G. Éric (Eds.), *Proceedings of the 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1273–1276).

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, *34*(4), 351–366. https://doi.org/10.1177/002383099103400404

Ang, J., Liu, Y., & Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)* (Vol. 1, pp. 1061–1064). https://doi.org/10.1109/ICASSP.2005.1415300

Austin, J. L. (1962). *How to do things with words*. Oxford University Press. https://doi.org/10.1093/acprof:oSo/9780198245537.001.0001

Azen, R., & Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, *34*(3), 319–347. https://doi.org/10.3102/1076998609332754

Beaver, D. I. (1997). Presupposition. In J. Van Benthem & A. Ter Meulen (Eds.), *Handbook of logic and language* (pp. 939–1008). Elsevier. https://doi.org/10.1016/B978-044481714-3/50022-9

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Bothe, C., Magg, S., Weber, C., & Wermter, S. (2018). Conversational analysis using utterance-level attention-based bidirectional recurrent neural networks. In B. Yegnanarayana, C. Chandra Sekhar, S. Narayanan, S. Umesh, S. R. Prasanna, H. A. Murthy, P. Rao, P. Alku, & P. K. Ghosh (Eds.), *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 996–1000). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-2527

Bothe, C., Weber, C., Magg, S., & Wermter, S. (2018). A context-based approach for dialogue act recognition using simple recurrent neural networks. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H.

Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)* (pp. 1952–1957). European Language Resources Association.

Boux, I. P., Margiotoudi, K., Dreyer, F. R., Tomasello, R., & Pulvermüller, F. (2023). Cognitive features of indirect speech acts. *Language, Cognition and Neuroscience*, *38*(1), 40–64. https://doi.org/10.1080/23273798.2022.2077396

Boux, I., Tomasello, R., Grisoni, L., & Pulvermüller, F. (2021). Brain signatures predict communicative function of speech production in interaction. *Cortex*, *135*, 127–145. https://doi.org/10.1016/j.cortex.2020.11.008

Brychcín, T., & Král, P. (2017). Unsupervised dialogue act induction using Gaussian mixtures. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 485–490).

Bucciarelli, M., Colle, L., & Bara, B. G. (2003). How children comprehend speech acts and communicative gestures. *Journal of Pragmatics*, *35*(2), 207–241. https://doi.org/10.1016/S0378-2166(02)00099-1

Bunt, H. (1994). Context and dialogue control. *Think Quarterly*, *3*(1), 19–31.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13–31.

Cerisara, C., Král, P., & Lenc, L. (2018). On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, *47*, 175–193. https://doi.org/10.1016/j.csl.2017.07.009

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. https://doi.org/10.1016/S1364-6613(02)00005-0

Chen, Z., Yang, R., Zhao, Z., Cai, D., & He, X. (2018). Dialogue act recognition via CRF-attentive structured network. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 225–234). https://doi.org/10.1145/3209978.3209997

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Colombo, P., Chapuis, E., Manica, M., Vignon, E., Varni, G., & Clavel, C. (2020). Guiding attention in sequence-to-sequence models for dialogue act prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 7594–7601. https://doi.org/10.1609/aaai.v34i05.6259

Dai, Z., Fu, J., Qile, Z., Cui, H., Li, X., & Qi, Y. (2020). Local contextual attention with hierarchical structure for dialogue act recognition. *arXiv preprint arXiv:2003.06044*. https://doi.org/10.48550/arXiv.2003.06044

Di Eugenio, B., Xie, Z., & Serafin, R. (2010). Dialogue act classification, higher order dialogue structure, and instance-based learning. *Dialogue & Discourse*, *1*(2), 1–24. https://doi.org/10.5087/dad.2010.002

Domaneschi, F., Passarelli, M., & Chiorri, C. (2017). Facial expressions and speech acts: Experimental evidences on the role of the upper face as an illocutionary force indicating device in language comprehension. *Cognitive Processing*, *18*(3), 285–306. https://doi.org/10.1007/s10339-017-0809-6

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, *76*(4), 444–463. https://doi.org/10.1086/648111

Duran, N., & Battle, S. (2018). Probabilistic word association for dialogue act classification with recurrent neural networks. In E. Pimenidis & C. Jayne (Eds.), *Proceedings of the 19th International Conference on Engineering Applications of Neural Networks (EANN)* (pp. 229–239). Springer. https://doi.org/10.1007/978-3-319-98204-5_19

Duran, N., Battle, S., & Smith, J. (2023). Sentence encoding for dialogue act classification. *Natural Language Engineering*, *29*(3), 794–823. https://doi.org/10.1017/S1351324921000310

Enrici, I., Adenzato, M., Cappa, S., Bara, B. G., & Tettamanti, M. (2011). Intention processing in communication: A common brain network for language and gestures. *Journal of Cognitive Neuroscience*, *23*(9), 2415–2431. https://doi.org/10.1162/jocn.2010.21594

Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, *25*(3), 207–237. https://doi.org/10.1080/09296174.2017.1366095

Garner, P. N., Browning, S. R., Moore, R. K., & Russell, M. J. (1996). A theory of word frequencies and its application to dialogue move recognition. In H. Bunnell & W. Idsardi (Eds.), *Proceeding of the 4th Inter-*

national Conference on Spoken Language Processing (ICSLP'96) (Vol. 3, pp. 1880–1883). IEEE. https://doi.org/10.1109/ICSLP.1996.607999

Gibbs, R. W. (1983). Do people always process the literal meanings of indirect requests? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 524–533. https://doi.org/10.1037/0278-7393.9.3.524

Gisladottir, R. S., Bögels, S., & Levinson, S. C. (2018). Oscillatory brain responses reflect anticipation during comprehension of speech acts in spoken dialog. *Frontiers in Human Neuroscience*, *12*, 34. https://doi.org/10.3389/fnhum.2018.00034

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS One*, *10*(3), e0120068. https://doi.org/10.1371/journal.pone.0120068

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)* (pp. 517–520). IEEE. https://doi.org/10.1109/ICASSP.1992.225858

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192. https://doi.org/10.3758/BF03195563

Grau, S., Sanchis, E., Castro, M. J., & Vilar, D. (2004). Dialogue act classification using a Bayesian approach. In *Proceedings of the 9th Conference Speech and Computer (SPECOM'2004)* (pp. 495–499). International Speech Communication Association.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

Haverkate, H. (1990). A speech act analysis of irony. *Journal of Pragmatics*, *14*(1), 77–109. https://doi.org/10.1016/0378-2166(90)90065-L

Hearst, M. A. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 257–323). Addison-Wesley.

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22. https://doi.org/10.1016/j.cognition.2019.03.005

Holtgraves, T. (2008). Automatic intention recognition in conversation processing. *Journal of Memory and Language*, *58*(3), 627–645. https://doi.org/10.1016/j.jml.2007.06.001

Hoque, M. E., Sorower, M. S., Yeasin, M., & Louwerse, M. M. (2007). What speech tells us about discourse: The role of prosodic and discourse features in speech act classification. In *Proceedings of the 2007 International Joint Conference on Neural Networks* (pp. 2999–3004). IEEE. https://doi.org/10.1109/IJCNN.2007.4371438

Hu, J., Gauthier, J., Qian, P. W., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Jang, Y.-M., Mallipeddi, R., Lee, S., Kwak, H.-W., & Lee, M. (2014). Human intention recognition based on eyeball movement pattern and pupil size variation. *Neurocomputing*, *128*, 421–432. https://doi.org/10.1016/j.neucom.2013.08.008

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)* (Vol. 1, pp. 364–367). IEEE. https://doi.org/10.1109/ICASSP.2003.1198793

Ji, G., & Bilmes, J. (2005). Dialog act tagging using graphical models. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)* (Vol. 5, pp. 33–36). IEEE. https://doi.org/10.1109/ICASSP.2005.1415043

Ji, G., & Bilmes, J. (2006). Backoff model training using partially observed data: Application to dialog act tagging. In R. C. Moore, J. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Comptutational Linguistics* (pp. 280–287). Association for Computational Linguistics. https://doi.org/10.3115/1220835.1220871

Ji, Y., Haffari, G., & Eisenstein, J. (2016). A latent variable recurrent neural network for discourse-driven language models. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 332–342). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1037

Jones, M. N. (2017). Developing cognitive theory by mining large-scale naturalistic data. In M. N. Jones (Ed.), *Big data in cognitive science* (pp. 1–12). Routledge.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual*. Institute of Cognitive Science, University of Colorado.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent convolutional neural networks for discourse compositionality. In A. Allauzen, H. Larochell, C. Manning, & R. Socher (Eds.), *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality* (pp. 119–126). Association for Computational Linguistics.

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1051–1056). Cognitive Science Society.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 719–724). Cognitive Science Society.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005

Khanpour, H., Guntakandla, N., & Nielsen, R. (2016). Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2012–2021).

Král, P., Cerisara, C., & Klečkova, J. (2006). Automatic dialog acts recognition based on sentence structure. In *Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (Vol. 1, pp. 61–64). IEEE. https://doi.org/10.1109/ICASSP.2006.1659957

Král, P., Pavelka, T., & Cerisara, C. (2008). Evaluation of dialogue act recognition approaches. In *Proceedings of the 2008 IEEE Workshop on Machine Learning for Signal Processing* (Vol. 29, pp. 492–497). https://doi.org/10.1109/MLSP.2008.4685529

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In S. A. McIlraith & K. Q. Weinberger (Eds.), *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (Vol. 32, pp. 3440–3447). AAAI Press. https://doi.org/10.5555/3504035.3504456

Lager, T., & Zinovjeva, N. (1999). Training a dialogue act tagger with the μ-TBL system. In A. Jönsson, N. Dahlbäck, A. Flycht-Eriksson, & P. Qvarfordt (Eds.), *Proceedings of the 3rd Swedish Symposium on Multimodal Communication*. Linköping University Natural Language Processing Laboratory (NLPLAB).

Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 515–520). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1062

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813313

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press. https://doi.org/10.7551/mitpress/5526.001.0001

Li, R., Lin, C., Collinson, M., Li, X., & Chen, G. (2019). A dual-attention hierarchical recurrent neural network for dialogue act classification. In M. Bansal & A. Villavicencio (Eds.), *Proceedings of the 23rd Conference on*

Computational Natural Language Learning (CoNLL) (pp. 383–392). Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1036

Li, W., & Wu, Y. (2016). Multi-level gated recurrent neural network for dialog act classification. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1970–1979).

Licea-Haquet, G. L., Reyes-Aguilar, A., Alcauter, S., & Giordano, M. (2021). The neural substrate of speech act recognition. *Neuroscience*, *471*, 102–114. https://doi.org/10.1016/j.neuroscience.2021.07.020

Linders, G. M., & Louwerse, M. M. (2023). Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, *30*, 77–101. https://doi.org/10.3758/s13423-022-02142-9

Liu, Y., Han, K., Tan, Z., & Lei, Y. (2017). Using context information for dialog act classification in DNN framework. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2170–2178). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1231

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589. https://doi.org/10.1111/tops.12349

Louwerse, M. M. (2021). *Keeping those words in mind: How language creates meaning*. Rowman & Littlefield.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426. https://doi.org/10.1111/j.1551-6709.2012.01269.x

Louwerse, M. M., & Crossley, S. A. (2006). Dialog act classification using n-gram algorithms. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference* (pp. 758–763). AAAI Press.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843–848).

Mac Cormac, E. R. (1985). *A cognitive theory of metaphor*. MIT Press.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, *3*(4), 303–328. https://doi.org/10.1016/0047-2727(74)90003-6

Milajevs, D., & Purver, M. (2014). Investigating the contribution of distributional semantic information for dialogue act classification. In A. Allauzen, R. Bernardi, E. Grefenstette, H. Larochelle, C. Manning, & S. W.-T. Yih (Eds.), *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)* (pp. 40–47). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1505

Mittlböck, M., & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, *15*(19), 1987–1997. https://doi.org/10.1002/(SICI)1097-0258(19961015)15:19%3C1987::AID-SIM318%3E3.0.CO;2-9

Noble, B., & Maraev, V. (2021). Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In S. Zarrieß, J. Bos, R. Van Noord, & L. Abzianidze (Eds.), *Proceedings of the 14th International Conference on Computational Semantics (IWCS)* (pp. 166–172). Association for Computational Linguistics.

Novielli, N., & Strapparava, C. (2009). Dialogue act recognition and the role of affect. In *Proceedings of the Doctoral Consortium at the 3rd International Conference on Affective Computing & Intelligent Interaction (ACII-2009)*. IEEE.

Ortega, D., & Vu, N. T. (2017). Neural-based context representation learning for dialog act classification. In K. Jokinen, M. Stede, D. DeVault, & A. Louis (Eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 247–252). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5530

Papalampidi, P., Iosif, E., & Potamianos, A. (2017). Dialogue act semantic representation and classification using recurrent neural networks. In V. Petukhova & Y. Tian (Eds.), *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue* (pp. 104–113). SEMDIAL.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). European Language Resources Association.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Raheja, V., & Tetreault, J. (2019). Dialogue act classification with context-aware self-attention. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3727–3733). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1373

Ravi, S., & Kozareva, Z. (2018). Self-governing neural networks for on-device short text classification. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 887–893). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1105

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Ribeiro, E., Ribeiro, R., & de Matos, D. M. (2015). The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*. https://doi.org/10.48550/arXiv.1506.00839

Ribeiro, E., Ribeiro, R., & de Matos, D. M. (2019a). A multilingual and multidomain study on dialog act recognition using character-level tokenization. *Information*, *10*(3), 94.

Ribeiro, E., Ribeiro, R., & de Matos, D. M. (2019b). Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, *66*, 861–899.

Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Prediction complements explanation in understanding the developing brain. *Nature Communications*, *9*(1), 589. https://doi.org/10.1038/s41467-018-02887-9

Ruytenbeek, N., Bergen, B., & Trott, S. (2023). Prosody and speech act interpretation: The case of French indirect requests. *Journal of French Language Studies*, *33*(1), 103–125. https://doi.org/10.1017/S0959269522000254

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*(4), 289–327. https://doi.org/10.1515/semi.1973.8.4.289

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, *5*(1), 1–23. https://doi.org/10.1017/S0047404500006837

Searle, J. R., & Vanderveken, D. (1985). *The foundations of illocutionary logic*. Cambridge University Press.

Shen, S.-S., & Lee, H.-Y. (2016). Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In N. Morgan, P. Georgiou, S. Narayanan, & F. Metze (Eds.), *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 2716–2720). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2016-1359

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) corpus. In C. Sidner & M. Strube (Eds.), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue* (pp. 97–100). Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems 33: 34th Conference on Neural Information Processing Systems 2020 (NeurIPS 2020)* (pp. 16857–16867).

Sridhar, V. K., Bangalore, S., & Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, *23*(4), 407–422. https://doi.org/10.1016/j.csl.2008.12.001

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, *25*, 701–721. https://doi.org/10.1023/A:1020867916902

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–373.

Surendran, D., & Levow, G. A. (2006). Dialog act tagging with support vector machines and hidden Markov models. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP'06)* (pp. 1950–1953).

Tinga, A. M., Kuperus, W., Carvalho, M. B., & Louwerse, M. M. (2019). Explanation versus prediction: Statistical differences in detecting fraudulent events do not necessarily have predictive power. In *Proceedings of the 41th Annual Conference of the Cognitive Science Society* (pp. 2975–2980). Cognitive Science Society.

Tomasello, R. (2023). Linguistic signs in action: The neuropragmatics of speech acts. *Brain and Language*, *236*, 105203. https://doi.org/10.1016/j.bandl.2022.105203

Tran, Q. H., Zukerman, I., & Haffari, G. (2017a). A generative attentional neural network model for dialogue act classification. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 524–529). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2083

Tran, Q. H., Zukerman, I., & Haffari, G. (2017b). Preserving distributional information in dialogue act classification. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2151–2156). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1

Tromp, J., Hagoort, P., & Meyer, A. S. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. *Quarterly Journal of Experimental Psychology*, *69*(6), 1093–1108. http://doi.org/10.1080/17470218.2015.1065282

Trott, S., Reed, S., Kaliblotzky, D., Ferreira, V., & Bergen, B. (2023). The role of prosody in disambiguating English indirect requests. *Language and Speech*, *66*(1), 118–142. https://doi.org/10.1177/00238309221087715

Verbree, D., Rienks, R., & Heylen, D. (2006). Dialogue-act tagging using smart feature selection; Results on multiple corpora. In M. Gilbert & H. Ney (Eds.), *Proceedings of the 2006 IEEE Spoken Language Technology Workshop* (pp. 70–73). IEEE. https://doi.org/10.1109/SLT.2006.326819

Wan, Y., Yan, W., Gao, J., Zhao, Z., Wu, J., & Philip, S. Y. (2018). Improved dynamic memory network for dialogue act classification with adversarial training. In N. Abe, H. Liu, C. Pu, X. Hu, N. Ahmed, M. Qiao, Y. Song, D. Kossmann, B. Liu, K. Lee, J. Tang, J. He, & J. Saltz (Eds.), *Proceedings of the 2018 IEEE International Conference on Big Data* (pp. 841–850). IEEE. https://doi.org/10.1109/BigData.2018.8622245

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://doi.org/10.1162/tacl_a_00321

Webb, N., Hepple, M., & Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding* (Vol. 4). AAAI Press.

Wierzbicka, A. (1987). *English speech act verbs: A semantic dictionary*. Academic Press.

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua*, *116*(10), 1722–1743. https://doi.org/10.1016/j.lingua.2006.05.001

Yano, Y., Tamura, A., Ninomiya, T., & Obayashi, H. (2021). Utterance position-aware dialogue act recognition. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1567–1574). INCOMA Ltd.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhao, T., & Kawahara, T. (2019). Effective incorporation of speaker information in utterance encoding in dialog. *arXiv preprint arXiv:1907.05599*. https://doi.org/10.48550/arXiv.1907.05599

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.